Supporting Information Appendix

for

Automated Vocal Analysis of Naturalistic Recordings from Children with Autism, Language Delay and Typical Development

D. Kimbrough Oller*, Partha Niyogi, Sharmistha Gray, Jeffrey A. Richards, Jill Gilkerson,

Dongxin Xu, Umit Yapanel, Steven F. Warren

*Correspondence to koller@memphis.edu

TABLE OF CONTENTS

SUPPORTING METHODS

Statistical analysis summary

Analyses to assess prediction of vocal development. Multiple linear regression (MLR) was the primary method used with the automated acoustic analysis to model vocal development. In the first step, SVI/SCU ratios for each of the 12 acoustic parameters for each day-long recording were regressed linearly against age in months at the time of each of their recordings for the typically developing children. The result was a normative model of vocal development as predicted by the acoustic parameters through entirely automated means (with no human intervention). In the second step, the SVI/SCU ratios for each of the recordings for the language delayed and autism samples were plotted on the basis of the normative MLR model, thus providing a comparative view of vocal development in the three groups as predicted by the automated acoustic parameter analysis for typically developing children (main text Figure 2b-g).

Analyses to model differentiation of groups. Classification predictions for children and recordings as typically developing, autistic or language delayed on the basis of the automated acoustic analysis were accomplished by a linear model, the parameters of which were estimated using linear discriminant analysis (LDA) and linear logistic regression (LLR) in separate analysis runs. Our primary method was based on the holdout method called **leave-one-out-cross-validation (LOOCV)** over the entire data set of 1486 recordings (802 typical, 333 language delayed, and 351 autistic) and 232 children (106 typical, 49 language delayed, and 77 autistic) with class predictions made by thresholding the output of the model on a held out data point (all the data from recordings of a single child) on each pass. This provided a jackknifed estimate of the prediction error in the standard way.

The use of a holdout method is critical in statistical modeling of this sort, because without it one runs the risk that findings will not generalize well to any new sample. LOOCV is a holdout method in which the statistical model is trained as many times as there are data points, in this case as many times as there were children in the dataset. In other words, LOOCV is the holdout method where the size of the heldout sample is one . To exemplify from the current study, in the first modeling pass of LOOCV, data on recordings from one child were held out while the model was trained on all other children's data. Then the left-out child's posterior probability (PP) was determined and plotted based on that first model. In the second modeling pass, a second child was randomly selected and the process was repeated, with the second child's PP being determined and plotted based on the second model. The process was repeated to determine PPs for all children, and in each case the child's outcome was based on a model where his/her own vocal characteristics had not been involved in the development of that model. This method limits the likelihood that idiosyncrasies of any individual child can significantly skew the outcome, and the data can thus be anticipated to generalize well to new samples.

The appropriateness of the LOOCV method was confirmed through analyses (both LDA and LLR) for a wide range of randomly determined holdout sample sizes and compositions using the entire data set. Figure S1 provides the results, illustrating the robustness of the LOOCV approach, outcomes for which always fell well within the range of outcomes based on the 9000 holdout tests, which themselves were subject to wide variability depending on peculiarities of any randomly selected training vs. testing set. Consequently our primary focus in the present work is on results from LOOCV, because that method proves to be relatively stable and to have high generalizability.

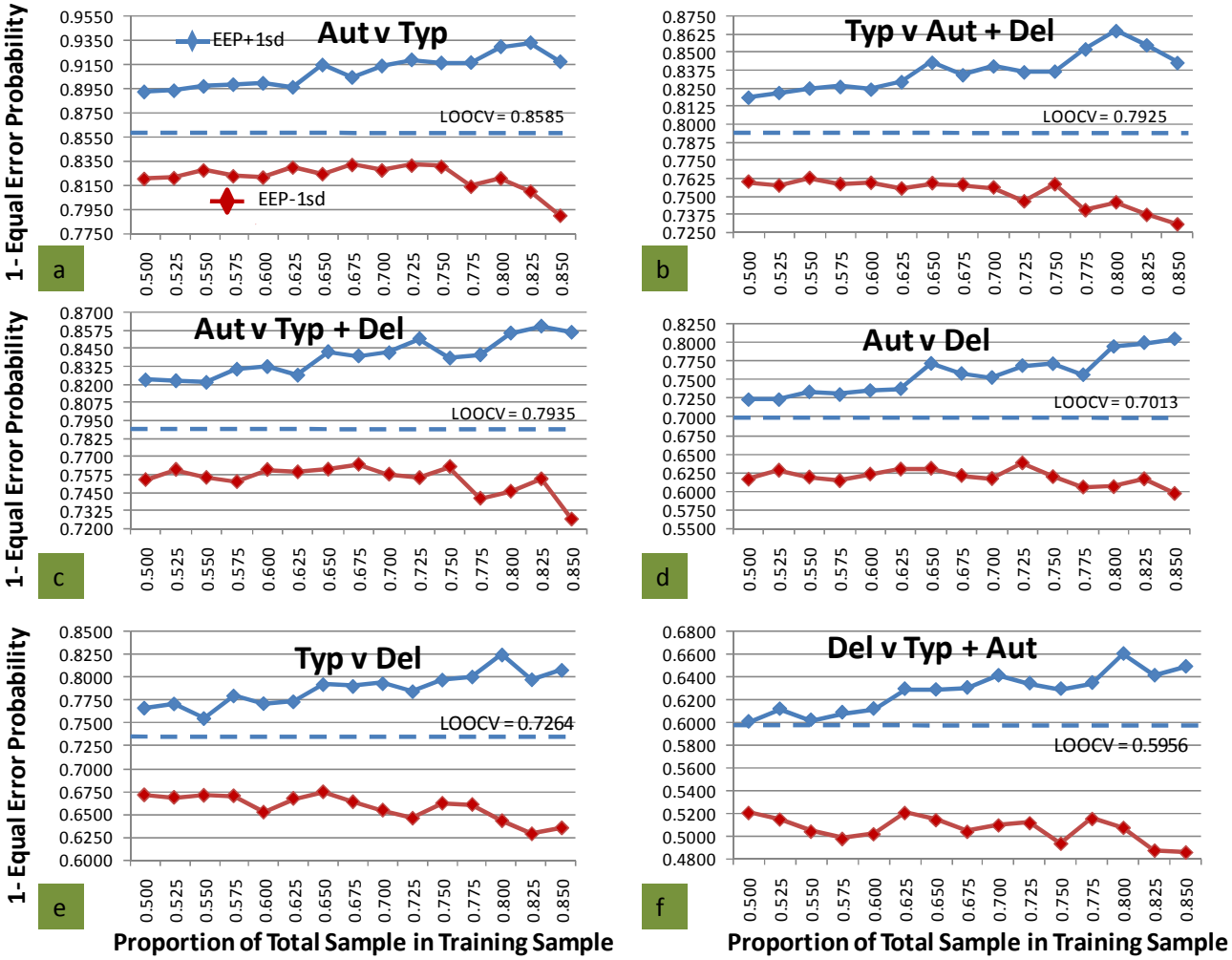

Figure S1. The six binary configurations of comparisons (a through f) among the groups illustrate that LOOCV analysis is well-motivated as a holdout procedure for analysis of the sort we conducted. In each panel (a through f), each x-axis location represents 100 training/testing runs (1500 runs per panel, 9000 in all) conducted in LDA. On each run a proportion (p) of the sample (ranging from 0.50 to 0.85) was selected for training as designated on the x-axes. LDA was used to model discriminability of the binary configuration (e.g., in panel a, the configuration **autism** sample **vs typical** sample), and the remainder of the samples (1 - p, ranging from 0.50 to 0.15) were used to test discriminability of groups based on the models. 100 such randomly selected runs yielded a range of outcomes for correct identification of children at equal sensitivity and specificity

(labeled "1 - equal error probability" on the y-axes). This range is represented with blue diamonds at one SD above the mean for the 100 runs at each x-axis location and with red diamonds at one SD below the mean. The means themselves are not plotted, but they can be determined by the reader by imagining a point midway between the blue and red points at each x-axis location. The dotted blue lines represent the value obtained for each configuration with LOOCV, for which of course there was only one value on each configuration. Note, for example, that the LOOCV sensitivity and specificity outcome of 0.8585 for Aut vs. Typ in panel a, falls in the midrange of the outcomes from the many randomly selected training/testing sets. For all six configurations, the LOOCV method yielded results within the two SD range displayed at every x-axis point representing different proportions of training and testing sets. It is especially important that the outcome for LOOCV appears conservative with regard to the configurations primarily focused on in this research, configurations a – c, falling near the mean value for the randomly selected tests with these configurations.

All Data Combined, LOOCV Comparisons (Ns: Typ = 106, Aut = 77, Del = 49)

a LDA	Sens/Spec	Chi Square	p	kappa
Aut v Typ	0.8585	92.92	5.46E-22	0.7118
Aut v Typ + Del	0.7935	73.78	8.74E-18	0.5576
Typ v Aut + Del	0.7925	79.29	5.37E-19	0.5842
Typ v Del	0.7264	28.27	1.06E-07	0.4171
Aut v Del	0.7013	19.57	9.69E-06	0.3904
Del v Typ + Aut	0.5956	5.72	0.016757	0.1360

b LLR	Sens/Spec	Chi Square	p	kappa
Aut v Typ	0.8585	92.92	5.46E-22	0.7118
Aut v Typ + Del	0.7922	73.10	1.23E-17	0.5550
Typ v Aut + Del	0.7925	78.92	6.46E-19	0.5828
Typ v Del	0.7143	25.25	5.04E-07	0.3934
Aut v Del	0.6939	18.15	2.04E-05	0.3758
Del v Typ + Aut	0.5792	3.91	0.047947	0.1114

All Groups Comparisons
Train Models on Phase 1 Data (Typ = 76, Aut = 34, Del = 28)
Then Test Models on Phase 2 Data (Typ = 30, Aut = 43, Del = 21)

c LDA	Sens/Spec	Chi Square	p	kappa
Aut v Typ	0.8667	38.68	4.99E-10	0.7270
Aut v Typ + Del	0.7255	19.01	1.3E-05	0.4492
Typ v Aut + Del	0.8	30.87	2.76E-08	0.5659
Typ v Del	0.7333	10.83	0.000998	0.4587
Aut v Del	0.6667	6.36	0.011693	0.3060
Del v Typ + Aut	0.5714	1.34	0.247282	0.1036

d LLR	Sens/Spec	Chi Square	p	kappa
Aut v Typ	0.8667	38.68	4.99E-10	0.7270
Aut v Typ + Del	0.7442	22.30	2.33E-06	0.4865
Typ v Aut + Del	0.8	30.87	2.76E-08	0.5659
Typ v Del	0.7143	9.13	0.002516	0.4208
Aut v Del	0.6512	5.22	0.022363	0.2765
Del v Typ + Aut	0.5714	1.34	0.247282	0.1036

Table S1. The LOOCV method shows remarkably similar outcomes for group discrimination under Linear Discriminant Analysis (LDA, panel a) and Linear Logistic Regression (LLR, panel b). In all six binary configurations of comparisons for panels a and b the discriminability was statistically reliable, and in 5 of the six, the reliability levels were very high (at least $p<10^{-4}$) (panel a, some data here are the same as in the main text Figure 4, but Chi Square and kappa values are added). When comparisons were conducted with a holdout method where Phase I data were used for training and Phase II data for testing (panels c and d), the outcomes were similar for group discrimination under both LDA (panel c, some data here are the same as in the main text Figure 4) and LLR (panel d). Statistical significance levels were lower in the bottom panels than the

top ones because the sample size for the Phase II test group was < 20% of the 232 children, but sensitivity/specificity values were similar.

LDA/LLR comparison and z-score conversion. To control for age effects, we used the distribution of the 12 acoustic parameters (ratio scores SVI/SCU) for the pooled data at each age interval and converted them to z-scores for each recording. The mean and SD for the typically developing recordings were used as the basis for the z-score computations for all three child groups. The intervals were windowed as follows: Each was centered at a month from 11-47 months; each window was four months wide, thus overlapping substantially with adjacent intervals. Each parameter was represented as an age-normalized z-score rather than a raw score in our LDA and LLR analyses.

The output of LLR can be directly interpreted as an estimate of the PP of autism. For the case of LDA, a further assumption was made that the data for each class had a Gaussian distribution and posterior probabilities were estimated under this assumption. Table S1 provides the data comparing outcomes of LLR and LDA as classification procedures in our research. As can be seen, the results were extremely similar for LDA and LLR, whether we examined the LOOCV approach or a different holdout approach where Phase I data were used for training and Phase II data were held out for testing. Consequently, in the main text of the article, we report outcomes on just one of the two methods, namely LDA.

The recording device

Basic characteristics. The LENA (Language ENvironment Analysis) recorder (to view the device and clothing, go to http://www.lenafoundation.org/ProSystem/Overview.aspxcan) can securely snapped into the chest pocket of specially designed clothing after it is set to record by an adult; it can be turned off only by holding the record button down for several seconds. When fully charged, the recorder can hold 16 hours of acoustic data (16 kHz sampling rate). It has a single microphone that remained 7-10 cm from the infant or child's mouth as long as it is snapped into the child's chest pocket. Given the close mouth-to-microphone distance, the recording quality is good in circumstances of low noise, but signal-to-noise ratio is affected negatively whenever there are interfering sounds in the environment, including sounds made if anyone touches the area on the clothing where the recorder is housed. The device has been involved in several previously published papers (45-48).

Versions of the recorder. Recording data (see next section) have been collected under three primary versions of the device since 2006. Tests have revealed no statistically reliable differences among outcomes for the basic utterance labeling (which indicates if a sound is the voice of a female adult, a male adult, the child wearing the recorder, television or radio, etc.) with the different recorder versions. The papers cited in the prior section (45-48) focused on analyses based on this basic labeling, and consequently the authors were empirically supported in employing data from all the recorder versions.

The analysis in the present work was different, however: Unlike in the prior studies, specialized acoustic analysis was done on sequences determined by the software to be speech-related child utterances (SCUs), and these SCUs were further segmented into roughly syllabic units (SVIs) by software not involved in any of the prior studies. In addition, none of the prior cited studies sought to model group discrimination on infrastructural acoustic parameters for infant sounds with LDA or LLR. We have found through comparison tests that the differing recorder characteristics across versions could indeed have affected some of the results in the discriminatory modeling that was done in the present study. To explain: Most of the children in the Phase I typically developing group had been recorded with the first or second recorder versions, but all the language delayed and autistic children had been recorded with the third. Thus it is possible that if we had used all the data from all recorder versions, some portion of the group discrimination achieved by LDA or LLR could have been due to recorder differences (presumably due to differing noise characteristics associated with recordings from the three recorder versions) rather than to vocalization differences across child groups. Consequently, the research reported here utilized recordings from the typically developing group only if they had all been made with the third recorder version, hereafter "the matched recorder" version, because that version was used in recording all data from the other child groups.

Participant groups, recording and assessment procedures.

***Typically developing sample*.** Naturalistic recordings with the LENA recorder were made during 2006-2008 by Infoture Inc. of Boulder, CO and are described by Gilkerson and Richards (49) and (45-47). The recordings are now the property of the LENA Foundation at the same location in Boulder, CO. This initial (Phase I) sample was collected in metropolitan Denver, including 328 typically developing, English-learning children, 2 months to 48 months of age. Families were matched to the US Census distribution for mother's educational level at each one-month interval of child age.

Prior to each recording, families who had given informed consent (approval by Essex Institutional Review Board, IRB) received a recorder plus an instruction sheet by overnight courier. After the recording, the device was returned to Boulder by courier for processing. Recordings were routinely made from the time children were dressed in the morning until bedtime. Children in this sample contributed multiple day-long recordings (averaging more than 5 per participant) at one-month intervals on different days of the week each month. Parents were compensated for participation.

All three recorder versions were utilized in this sampling, but a subset (N = 76) of the typically developing children in Phase I (the "longitudinal sample", followed for a longer period than other children in the samples) utilized the matched recorder version only. These are the children whose data are compared in the

present study with the autism and language delayed samples (who of course were also recorded with the matched recorder version).

For the present research, 712 recordings at 10-48 months from 76 of the typically developing children from Phase I of the LENA Foundation natural language sample were examined (see Table S2). The 712 recordings were selected for the present study because they fell within the age range of available recordings from the other child groups in addition to having been made with the matched recorder version.

Table S2	CHILDREN				RECORDINGS		
	Phase I	**Phase II**	**TOTAL**		**Phase I**	**Phase II**	**TOTAL**
Typically developing, N **Recording age range** **Recruitment**	76 10-48 mo. Denver Metro	30 18-37 mo. National on-line	**106**		712	90	**802**
Language delayed, N **Recording age range** **Recruitment**	28 10-40 mo. Denver Metro	21 22-44 mo. National on-line	**49**		270	63	**333**
Autism, N **Recording age range** **Recruitment**	34 16-48 National **not** on-line	43 24-48 National on-line	**77**		225	126	**351**
TOTAL	**138**	**94**	**232**		**1207**	**279**	**1486**

Table S2: Children and recordings involved in the present study

In the course of participation, parents periodically filled out questionnaire assessments. One was the LENA Developmental Snapshot, based on 52 questions compiled from a wide variety of existing standardized instruments. The Snapshot is an easily administered evaluation conducted periodically to assess communicative development of all the children. Details on Snapshot items and reliability/validation data can be found in Gilkerson and Richards (50). Of particular importance here are the high correlations obtained between Snapshot scores and scores on the most widely accepted early communication scales, the PLS-4 (51) (Snapshot score correlation with PLS-4 score on expressive language, r = 0.92; on receptive language, r = 0.93) and the REEL-3 (52) (expressive r = 0.96, receptive r = 0.96). An additional important evaluation administered for nearly all the children in all the groups (see Table S3), the CDI (Child Development Inventory) (53), offered general measures of language/communicative, cognitive and social development. A final parent-questionnaire assessment was the CBCL (Child Behavior Checklist) (54), providing information about socio-psychiatric factors including possible autistic symptoms (see below, **Analyses indicating appropriate characteristics of the participant groups**). Phase I typically developing children were brought into our laboratories in Boulder and evaluated during the course of their recordings by a staff speech-language pathologist who administered the REEL-3 to 72 of the children and the PLS-4 to all 76.

Table S3	Snapshot		CSBS		CDI		MCHAT		CBCL		ADOS		CARS	
	Ph I	Ph II	Ph I	Ph II	Ph I	Ph II	Ph I	Ph II	Ph I	Ph II	Ph I	Ph II	Ph I	Ph II
Typically developing	ALL	ALL	NONE	ALL	73	ALL	NONE	ALL	42	ALL	NONE	NONE	NONE	NONE
Language delayed	ALL	ALL	NONE	ALL	ALL	ALL	NONE	ALL	NONE	ALL	NONE	NONE	NONE	NONE
Autism	ALL	ALL	ALL	ALL	32	ALL	ALL	ALL	31	ALL	17	12	9	15

Table S3: Number of children for whom results of various assessments were obtained

In Phase II beginning in 2009, additional recordings were made with matched recorders for children from families living in a variety of locations around the USA in order to expand the database, to improve the potential for hold-out sample testing of the automated procedure (especially with randomly selected holdout samples for comparison with each other and with LOOCV), to enhance our ability to evaluate within-group variations in vocal patterns, and to provide the opportunity to collect data relevant to diagnosis based on parent questionnaire evaluations for all three groups of children. The Phase II participants were primarily recruited by internet advertisement. Parents enrolled by filling out an on-line form. This we refer to as the "national on-line" recruitment method (Table S2). 30 typically developing children were recorded based on recruitment in Phase II. The procedures for consent and recording were essentially the same as those of Phase I. However, only three recordings were made per child in Phase II, all occurring during a 7-10 day period. Also in Phase II, additional parent questionnaire assessments of language development and social status of the children were obtained (see Table S3). These assessments included the MCHAT (Modified Checklist for Autism in Toddlers) (55) and the CSBS (Communication and Symbolic Behavior Scales) (56), which were not obtained in Phase I for either the typically developing or language delayed samples. These two evaluations were conducted for children in all three groups during Phase II (see Table S3).

The Phase I and Phase II typically developing samples were extremely similar in chronological age (Phase I M = 28.5 months, Phase II M = 27.3) and in general characteristics as indicated by the evaluations. There were no statistically significant differences between scores for children in the two phases for language or general development level as indicated by the Snapshot (either Developmental Age, Standard Score or Developmental Quotient) or the CDI (either general development or expressive or comprehension in language).

Autism sample. In Phase I, national recruitment (but not by the on-line method) for children who had been formally diagnosed with autism yielded 34 qualifying and consenting families. In Phase II, an additional 43 were recruited using the national on-line method. The entire autism sample appears to be typical of children commonly diagnosed with ASDs in early childhood (see below **Analyses indicating expected characteristics of the participant groups).** Children were selected to show low language skills compared to age-matched

typically developing children and to show other symptoms of autism which could include stereotypic movements, a tendency to avoid eye-contact, and so on. The selection procedure excluded Asperger syndrome. Parents were asked to send copies of comprehensive evaluations indicating a formal diagnosis of autism from physicians, psychologists, and/or other professionals as a condition of participation. The documentation of autism diagnosis was extensive. The written reports from diagnostic workups from health professionals and parent reports regarding these workups indicated varying degrees of severity. Three children from Phase I and two from Phase II were included in the autism sample even though their formal diagnosis was PDD (Pervasive Developmental Disorder), the diagnosis that is often applied when a child shows symptomatology consistent with autism, but is younger than the usually assumed age threshold (36 months) for formal diagnosis of classic autism.

In Phase I, the families of the 34 children with autism who participated were instructed to record their infants once weekly eight times across seven weeks. Recordings were staggered, according to instruction from project staff, to ensure that weekend and weekday samples were represented for all children. The analyses provided here for Phase I were based on 225 recordings (2 to 8 per child), the ones available when the analysis was conducted, all with the matched recorder version. In addition the Snapshot, CSBS, CDI, MCHAT and CBCL were obtained for nearly all children in the autism sample (see Table S3).

In Phase II, the 43 additional children with autism were recorded according to the same schedule as with the Phase II typically developing sample. In addition the same parent evaluations as for the Phase II typically developing sample were administered (Table S3).

The Phase I and Phase II autism samples differed in chronological age (Phase I M = 33.6 months, Phase II M = 37.8), but like the typically developing sample, were otherwise similar across phases in general characteristics as indicated by the evaluations. There were no statistically significant differences between scores for children in the two phases for language or general development level as indicated by the Snapshot (Standard Score or Developmental Quotient) or the CDI (neither general development, expressive language, nor comprehension of language), although, consistent with the chronological age differences, raw scores and developmental ages were somewhat higher in Phase II children.

Language delayed sample. For the language-delayed sample, 28 children were recruited in Phase I from the Denver metropolitan area based on a prior formal diagnosis by speech-language pathologists or pediatricians of communication delay, not designated as being associated with autism. The sample had been selected to be age-matched with the typically developing sample, which originally included children through three years of age, but was later extended to 48 months, after the language-delayed group had already been recruited and largely recorded; hence recordings from the language delayed sample were not obtained beyond age 40 months in Phase I. Procedures for recruitment and intake were similar to those for the autism sample.

The children had all been diagnosed by speech-language pathologists or pediatricians as having speech or language delays or, in the cases of the youngest children, developmental delays involving or expected to involve communication or vocalization. The specific diagnoses were various, including many who were simply characterized as having speech-language delays, and several with mixed diagnoses including two with apraxia/dyspraxia, four with sensory integration disorder/central auditory processing disorder, and three designated as failure to thrive. One infant had Down syndrome.

The children in the language-delayed samples of Phase I were also evaluated by a certified speech-language pathologist in the Boulder laboratories twice, once in the first month and once in the sixth month of the sampling. Both the REEL-3 and the PLS-4 (51, 52) were administered to each of the children with language delay in Phase I, and additional development evaluation notes were compiled based on the clinician's impressions. On the basis of this evaluation, four of the children with language delay were designated (unexpectedly) as having "autistic characteristics". These children are considered individually in results below (**Individual children and subgroups of particular interest in the group discrimination analyses**). For all children in the language delayed sample in Phase I, we also obtained the Snapshot and the CDI.

Recordings for children with language delay in Phase I were conducted monthly over a six-month period, with the exception that three recordings were made in the second and sixth months for a total of 10 per child. Otherwise the protocol was the same as for the longitudinal typically developing sample.

In Phase II, 21 additional children with language delay were recruited through the national on-line method, and recordings were conducted according to the same schedule as with the Phase II typically developing and autism samples. These children included a mixture of presumed etiologies similar to those of the Phase I language delayed children with the exception that the on-line recruitment procedure had yielded a larger proportion (7/21 as opposed to 2/28) of children designated as having apraxia. The same parent evaluations as for the Phase II typically developing and autism samples were administered (Table S3).

The Phase II language delay sample was statistically significantly older (evaluated by ANOVA with Posthoc Tukey's t-tests) than the Phase I sample (Phase I $M = 26.7$ months, Phase II $M = 32.0$, $p < .05$). In spite of this fact, language levels in Phase II were significantly lower ($p < 0.01$) as suggested by the Snapshot Developmental Quotient (Phase I $M = 75.8$; Phase II $M = 57.3$) and Standard Score (Phase I $M = -1.37$; Phase II $M = -2.40$). Importantly, in Phase II, language delay children were recruited specifically to show low scores on the Snapshot. However, the Snapshot Developmental Age for children from the two Phases was not reliably different, presumably because the fact that the Phase II sample was chronologically older tended to offset their lower language levels. In contrast, CDI results showed slightly higher scores for Phase II, but they were not statistically significant.

Summary of selection criteria for the participant groups. For both Phase I and Phase II, the following selection criteria were applied:

1. For typically developing children:
    a. No indication of developmental disorder
    b. English language in the home
    c. Age ≤ 48 months
2. For children with language delay:
    a. Parent report that a diagnosis of language delay had been given by a speech-language pathologist
    b. English language in the home
    c. No indication of autism in prior diagnosis
    d. Age ≤ 48 months
3. For children with autism:
    a. Parent report that a diagnosis of ASD had been given by a qualified professional
    b. Written diagnostic documentation supplied by parents to our staff from the professional(s) who had evaluated the child
    c. Asperger syndrome excluded
    d. English language in the home
    e. Age ≤ 48 months

For Phase II, the following *additional* selection criteria were applied:

4. For typically developing children:
    a. Age 18-36 months
    b. Not failing the MCHAT on either scoring option (not autistic)
    c. Not < 80 and not > 110 on the Snapshot (midrange language levels)
    d. No sibling with developmental delay diagnosis
    e. No sibling with autism
    f. No other symptoms of autism on intake questionnaire (frequently repeated motions, lack of eye contact)
5. For children with language delay:
    a. Not failing the MCHAT on either scoring option (not autistic)
    b. Written diagnostic documentation supplied by parents to our staff from the professional(s) who had given the diagnosis of language delay to the child
    c. At least 1.5 SD below the Standard Score mean on the Snapshot (clearly low language level)
    d. No sibling with autism
    e. No other symptoms of autism on intake questionnaire (frequently repeated motions, lack of eye contact)
6. For children with autism:
    a. Fail MCHAT on at least one of two scoring options

Ethnicity was not a factor in subject selection at any time for this study, although non-English-speaking families were excluded. Table S4 presents the data on ethnicity, indicating that the predominant ethnic group was Caucasian/not Hispanic in all three child groups (typically developing M = 81%, language delayed M = 84%, autism M = 83%).

Table S4	Typically developing		Language delay		Autism		
	Phase I	Phase II	Phase I	Phase II	Phase I	Phase II	TOTAL
African-American	5	2	0	0	0	2	**9**
Asian	0	0	2	1	1	2	**6**
Hispanic	7	1	2	2	4	1	**17**
Native American	1	0	0	0	0	1	**2**
Caucasian/not Hispanic	59	27	23	18	27	37	**191**
Other	4	0	1	0	2	0	**7**
TOTAL	**76**	**30**	**28**	**21**	**34**	**43**	**232**

Table S4. Ethnicity breakdown for the three participant groups.

Analyses indicating appropriate characteristics of the participant groups. Our participant families were volunteers meeting the criteria indicated above. We evaluated the possibility that the samples could have been skewed by self-selection in ways that might have distorted the research outcomes. In particular we evaluated the extent to which the samples showed the usual characteristics associated with their diagnosis or lack of diagnosis as indicated by the parent-questionnaire evaluations. Figure 1b and 1c in the Main Text show, for example, that language level on the Snapshot was low for the language delayed sample and even lower for the autism sample, as expected. Taking the standard error data from the Figure into account, the language level differences among the groups were extremely large, and ANOVA with Posthoc Tukey's t-tests indicated they were highly statistically significant ($p<10^{-8}$).

Similarly, scores were available for both the CDI on both expression and comprehension of language for 87% of the 232 children. Both raw scores and age equivalent scores showed that as expected the typically developing children's language levels (age equivalent expressive M = 31.9 months, SE = 1.3; comprehension M = 28.9, SE = 1.2) were robustly higher ($p < 0.0001$) than for either of the other groups, who showed similar outcomes on this measure (language delayed expressive M = 22.5 months, SE = 0.9; comprehension M = 22.6, SE = 1.0: autism expressive M = 23.6 months, SE = 1.0, comprehension M = 21.6, SE = 1.0). The age equivalent score difference is particularly revealing given that the mean real age of the typically developing sample was considerably lower than that of the autism sample and narrowly lower than that of the language delay sample.

Figure S2 shows group differences, as expected, on socio-psychiatric measures obtained through the CBCL. Very robust differences were found between the autism and typically developing samples on all measures. Also the language delayed sample fell between the typically developing and autism samples on 11 of the 12 measures, and differences between the language delay sample and the other groups were strong in more than half the cases (see CIs in the figures). Especially notable were the very high values for the autism sample on the two factors deemed most indicative of autism (Pervasive DD and Withdrawal), both falling above levels designated for the clinically disordered range (dotted lines).

The current autism sample (N = 74, including all the children for whom the relevant test data were available, mean age = 36 months) was very similar on the CBCL measures to that of an independent autism sample reported by Sikora, et al. (57) (N = 50, mean age = 50 months, see Figure S3). Again both samples fell above thresholds for clinically significant values on Pervasive DD and Withdrawal.

Additional analysis using CSBS (56, 58) data, available for typically developing and language delayed children in Phase II and for all children in the autism sample, showed a similar pattern to that for the CBCL. Composite score on social factors, presumed to be particularly indicative of autism, were dramatically different for the child groups, with the autism sample again falling beyond a clinically significant threshold. The groups were also very strongly differentiated on the symbolic composite score. On the speech composite, the two language-disordered groups scored very similarly but differed dramatically from the typically developing sample. The younger sample of children from Wetherby et al. (56, 58) (for all three groups *M* = 21 months; our study typical *M* = 28 months, language delay *M* = 29 months, autism *M* = 36 months) also showed stark differences between autism and typically developing children on all three composite scores, see Figure S4.

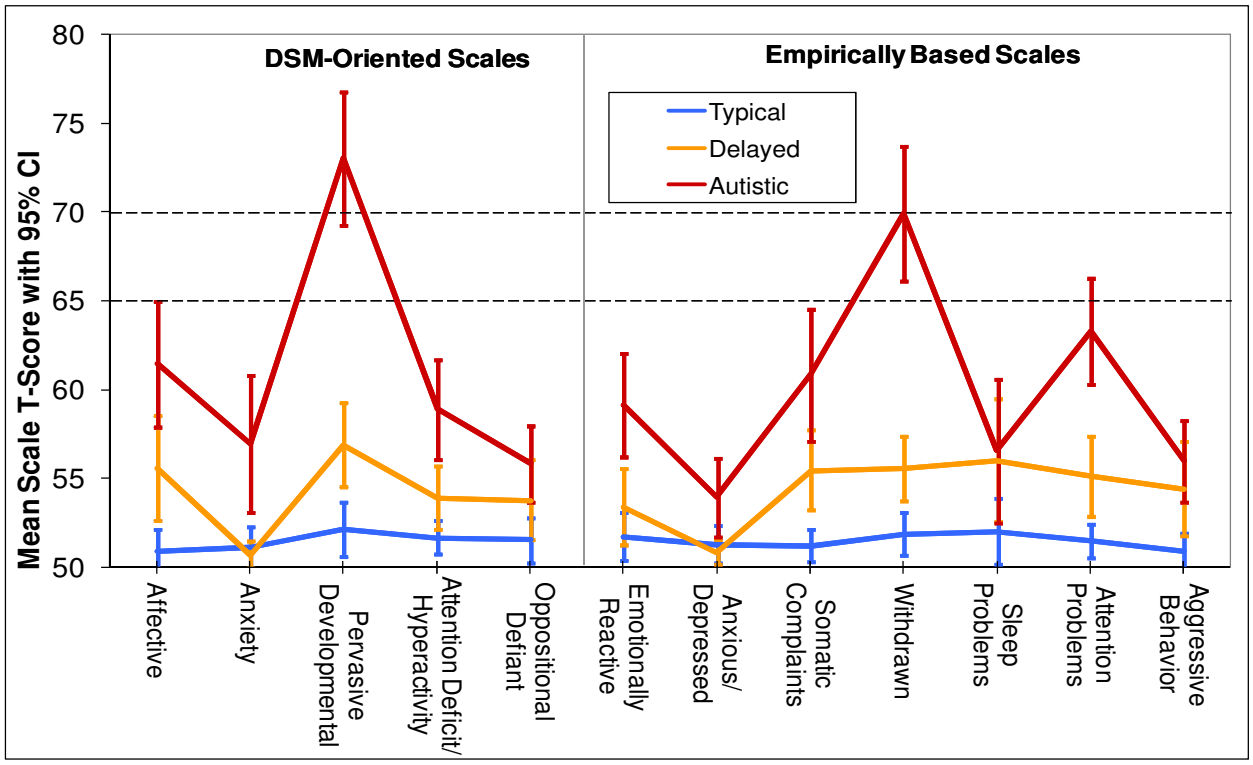

Figure S2: Group differentiation for the current sample (Phase II) on diagnostic features of the CBCL. These data are based on 94 children, 285 recordings collected in Phase II, which included administration for all participants of a parent questionnaire designed to assess social and psychiatric parameters, the Child Behavior CheckList (CBCL, (54)). The CBCL consists of two groupings of characteristics (DSM and Empirically-based). Children with autism were rated statistically reliably higher on all the characteristics than the typically

developing children, but especially notable were extremely high values for children with autism on primary characteristics known to be associated with autism, namely pervasive developmental problems and withdrawal, where their ratings were in the clinically disordered range (dotted lines, 65 = clinical threshold, 70 = severe rating).

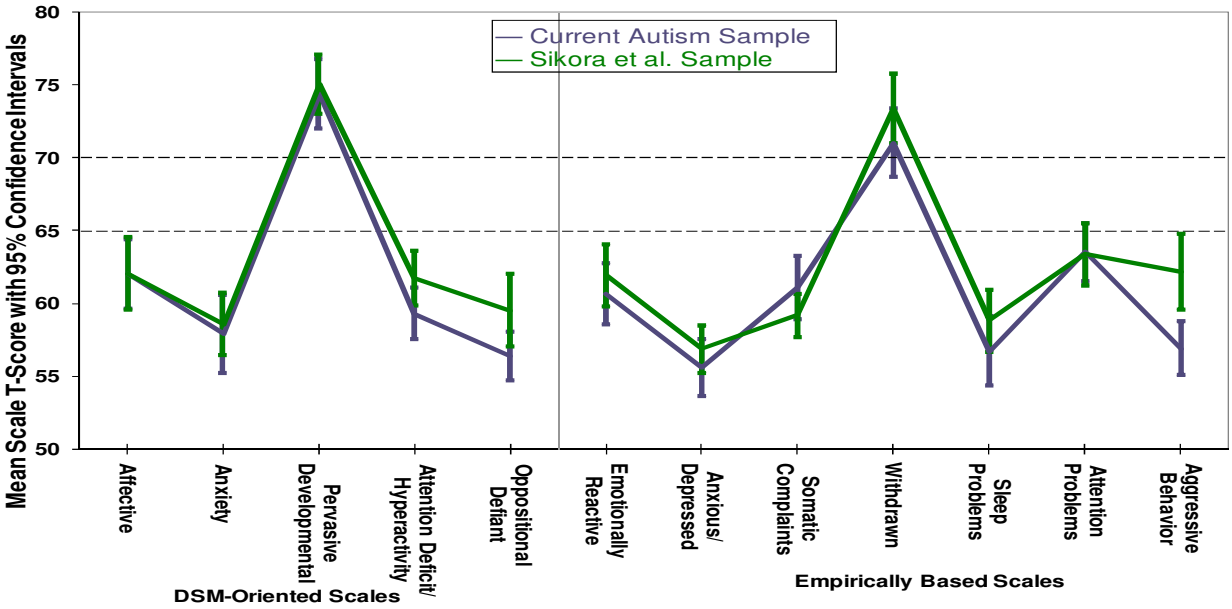

Figure S3: Comparison of current autism sample with that of a published autism sample on CBCL measures

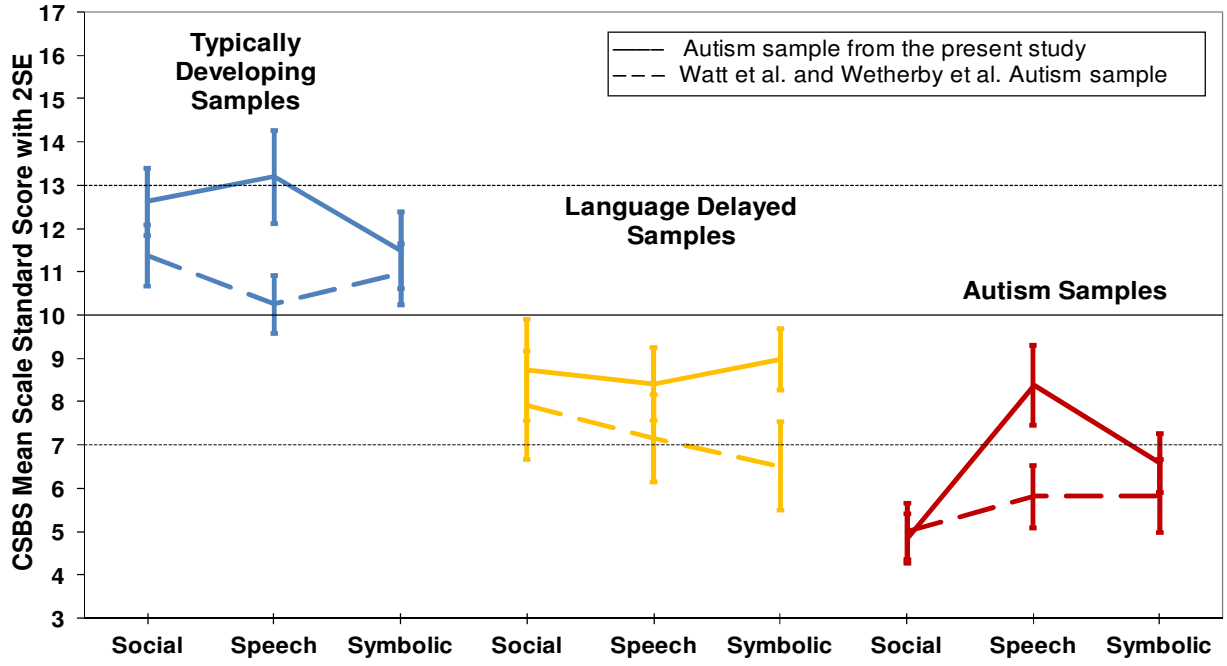

Figure S4. Data on Social, Speech and Symbolic Scales of the CSBS (56, 58) for the current samples and for somewhat younger children from the Wetherby et al. sample.

Automated analysis algorithms

Targeted acoustic sequences in the recordings. The goal of the automated analysis was to locate key child utterances (i.e., vocalizations produced by the infant or child wearing the recording device, as opposed to any other child in the environment) that could be deemed to be speech or indicators of an emerging capacity for speech. Thus acoustic events resembling speech or representing presumably voluntary vocalizations that are thought to be precursors to speech (babbling, cooing, squealing, growling, etc.) were included in the analysis, while such sounds as cries and vegetative noises, presumed to be relatively involuntary, were excluded. Also child utterances were grouped into clusters or phrases consisting of key child utterances as they occurred in the stream of vocalization and conversational interchange with other speakers. Five basic steps of analysis were used to locate the acoustic signals for analysis, and in step 6 the acoustic analysis occurred.

Step 1: Utterance cluster (CUC) location. The first step automatically located and labeled acoustic signals corresponding to the key child voice as child utterance clusters (CUCs). (The term CUC is used in this paper although the LENA software actually assigns labels CHN, for "child near", that is, highly audible and CHF, for "child faint or fuzzy", to the vocalizations in question). CUCs were defined to have at least 600 ms duration and to consist of periods identified as pertaining to the voice of the key child, while not being interrupted by utterances of any other speaker (labeled as male adult, female adult, or other child) and also not being interrupted by silence or noise of more than 800 ms as illustrated in Figure S5. The method was designed to locate relatively continuous periods of infant or child vocalization that were either produced as key child contributions to vocal interactions or as child monologues. The identification was based on a maximum likelihood algorithm using Gaussian mixture models that had been trained to recognize the key child voice and other types of sounds.

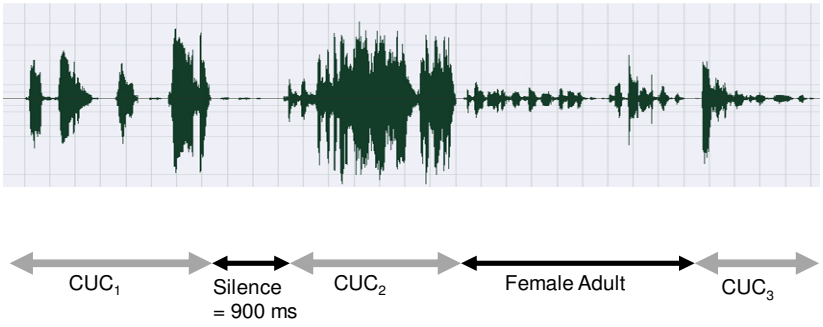

Figure S5. A waveform example including a key child and a female adult voice as identified by the analysis algorithm. By definition, child utterance clusters (CUCs) are bounded by not less than 800 ms of silence or sound not produced by the key child. A silence of 900 ms and a female adult voice of substantially greater duration separate the CUCs in the figure. See Supporting text for details of the procedure of CUC identification.

The limiting values of 600 and 800 ms for duration of CUCs and interruptions of them by other sounds or silences respectively were selected primarily on the basis of the empirical performance of the system in classifying utterances in such a way as to match auditorily-based judgments from human transcribers – agreement with the auditory codings proved best when these limiting values for automatic labeling were used. The values are also, however, theoretically justifiable: 600 ms can be said to represent a duration for an utterance consisting of one long syllable or two syllables (a typical metrical foot) of usual duration in adult or child speech (59, 60). Consequently the 600 ms constraint focused the analysis on CUCs with a relatively high minimum level of prominence. The 800 ms minimum value for interruptions between CUCs was similarly theoretically based; pauses between elements of a conversation are commonly 800 ms or longer (61-63).

To achieve this automatic recognition of CUCs and other sounds, 230 hours of recording from the dataset of infant and child recordings from 2-48 months were first coded auditorily by human transcribers to identify each vocalization sequence perceived as being produced by the key child, another child, a female adult, a male adult, or voices from television, stereo, radio or electronic toys. Other noises and silence, as well as overlapping sounds of any of the prior categories were also coded, so that the training recordings were labeled exhaustively. The Gaussian mixture models were trained to mimic the auditory codings, exhaustively labeling the recordings in terms of the same categories.

Step 2: Locating utterances (CUs) within utterance clusters (CUCs). A CUC could contain one or more child utterances (CUs), not overlapping with or significantly interrupted by any of the other vocal categories, see Figure S6. The division of CUCs into CUs was accomplished in the second step of the automatic analysis, wherein periods of high energy within CUCs were identified (although sometimes there was just one high energy period in a CUC). The beginning of the first CU in a CUC was automatically labeled when the acoustic energy level first rose to 90% above baseline for at least 50 ms and ended when it fell to less than 10% above baseline for at least 300 ms.

The term "utterance" is drawn from the literature in infant vocalizations and child phonology where it is typically defined by a "breath-group" criterion (64, 65). The idea is that vocalization occurs on expiration, and each time an inspiration occurs (or each time there is a break in sound long enough for inspiration to occur, nominally 300 ms), listeners perceive a break between vocal events. Vocalization occurring during expiration can of course be broken up by brief articulatory events (consonantal-like closures of the supraglottal vocal tract); a combination of expiratory events and any number of these brief articulatory breaks without an inspiration is termed a "breath-group" or "utterance".

Child utterances (CUs) were thus defined to consist of breath-groups of child vocalization within CUCs separated by not more than 300 ms of other sound or silence. The value of 300 ms was chosen because it has been shown empirically to fall at the high end of the distribution for silences or for low energy periods

(corresponding usually to consonantal closures) occurring within an utterance (63, 66). Further, 300 ms can be thought of as a nominal average duration for a stressed syllable (the minimal utterance) in mature speech (67).

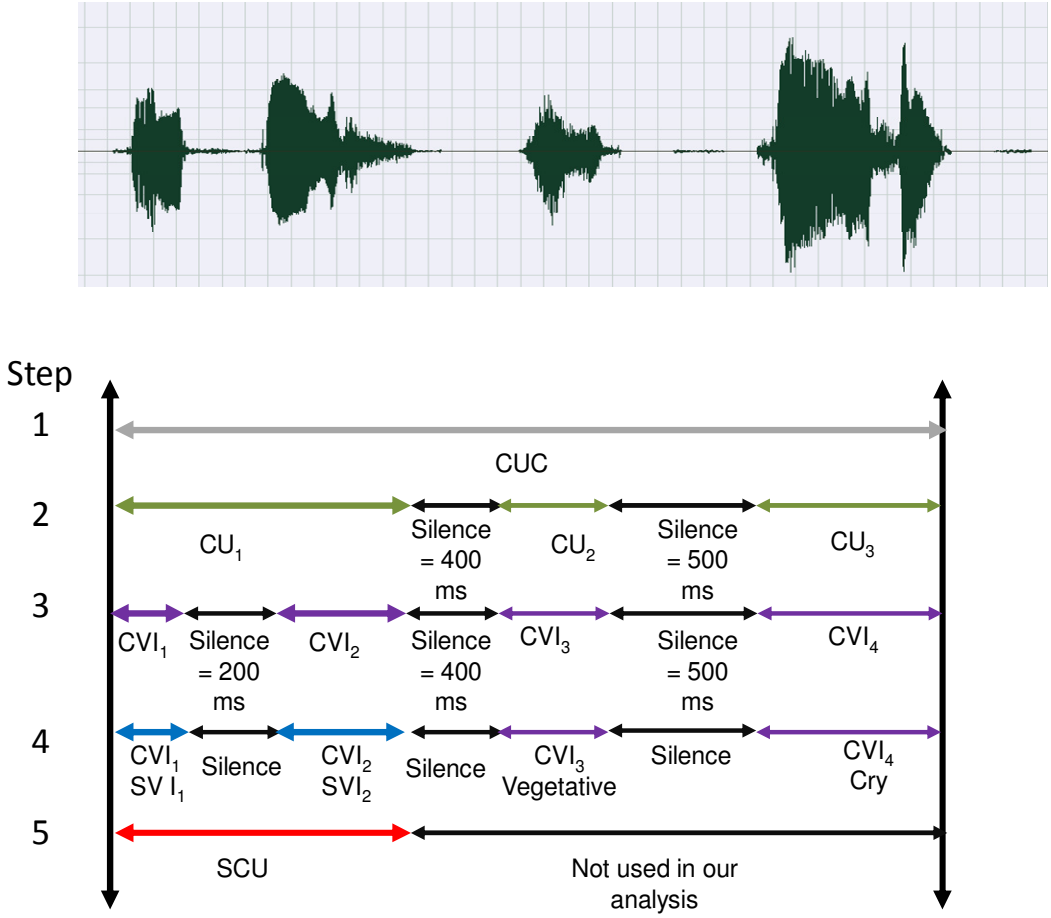

Figure S6. Waveform example showing that a CUC can consist of one or more child utterances (CUs). The first CUC from Figure S5 is expanded here to show the three child utterances (CUs) of which the cluster consisted. CUs are by definition bounded by not less than 300 ms of other sound or silence. The silence of 200 ms within CU_1 is a consonantal closure within the utterance. The figure also illustrates the five steps of the identification of segments. In the first two steps CUCs and the CUs of which the CUCs are composed are identified. In step 3 CVIs (child vocal islands), corresponding roughly to syllables within CUs, are located. In step 4 CVIs are further categorized as either Speech-related Vocal Islands (SVIs), or as vegetative sounds or cries. In step 5 any consecutive sequence of SVIs within any CU is further classified as a Speech-related Child Utterance (SCU). The SVIs (**blue arrow intervals** in the figure) are the focus of the acoustic analysis, and the primary measures addressed in this work represent ratios: The number of SVIs classified as "plus" within each recording on each parameter divided by the number of SCUs (**red arrow intervals** in the figure) within that recording. See Supporting text for details of the procedure of identification of utterances and SVIs at all five steps.

Step 3: Identification of child vocal islands (CVIs) within CUCs. In step 3 of the automated analysis, as indicated in Figure S4, "islands" of high energy were identified within CUs (often there was just one such island in a CU). A key child vocal island (CVI) was automatically labeled when the acoustic energy level rose

to 90% above baseline for at least 50 ms and ended when it fell to less than 10% above baseline for at least 50 ms, but not more than 300 ms, because as required by the utterance criterion, a CU boundary would have been inserted by the algorithm after 300 ms. In general CVIs correspond to syllables with very strong differentiations of acoustic energy level between nuclei (or vowels) and margins (or consonants).

Step 4: Classification of speech-related child vocal islands (SVIs). In the fourth step (see Figure S6), each key child vocal island or CVI was automatically assigned to one of three categories: (a) cry, (b) "vegetative sounds", consisting of such sounds as sneezing and coughing as well as laughter, and (c) speech-related vocal islands (SVIs), using another maximum likelihood algorithm based on Gaussian mixture models. SVIs included prespeech vocalizations of children such as cooing and babbling as well as real speech (In some of the prior writings regarding LENA analyses, the term "meaningful speech" has been used to refer to speech-related vocalizations. The present work adopts the term "speech-related" to ensure the understanding that both speech and many prespeech vocalizations such as cooing and babbling are understood to be included under the term. The term "SCU count" corresponds to the formal label "vocalization count" in the LENA software.).

The models corresponding to these categories had been developed through training with 23 hours consisting exclusively of waveforms that had previously been auditorily coded by human transcribers as (a), (b), or (c). There were 8 hours of (a), 4 of (b), and 11 of (c). The training materials had been drawn from 223 days of recording from the typically developing sample evenly distributed across the age range of 2 to 41 months. Gaussian mixture models were developed using the training materials for each of the three categories and for each age in months. Hence there were 3 x 40 or 120 Gaussian mixture models to be applied to the test sets from the three groups. This three-fold automated classification afforded segregation of the SVIs from other infant and child sounds (cries and vegetative sounds) that are arguably less relevant to the development of speech (68-70).

Step 5: Grouping of speech-related vocal islands into speech-related child utterances. In the fifth step (Figure S7), *consecutive* child vocal islands within any CU, all of which had been categorized as SVIs, were grouped into speech-related child utterances (SCUs) where (in accord with the utterance criteria specified in step 2 of the analysis) no interruption of the sequence by silence, cry, vegetative sounds, or any other voice could exceed 300 ms.

The logic of the analysis sequence allowed CVIs consisting of cries or vegetative sounds to occur within CUs, but cries and vegetative sounds were not included within SCUs. Thus a single CU could be broken up into multiple SCUs by interruptions between SVIs consisting of any consecutive sequence of algorithm-identified cries or vegetative sounds plus other periods not identified as the key child's voice. If such an interruption between SVIs within a CU exceeded 300 ms, a boundary was automatically established between two SCUs.

An example is provided in Figure S7, where the second CU includes four SVIs, broken up into two SCUs by a cry. To summarize the criteria for identification of SCUs, if a low energy period (a break between islands) was greater than 50 ms and less than 300 ms, it was treated as a within-vocalization consonantal (articulatory) event, separating two SVIs within an SCU, but if the break or any combination of a break and a cry or vegetative sound exceeded 300 ms, it was treated as a boundary between two SCUs. SVIs were thus grouped in such a way that they could contain a single syllable or a series of syllables where energy did not fall below the minimum of 10% above baseline consistent with the criterion defining the notion island. It is important to recognize that nasal or glide consonants, for example, typically do not fall below this energy baseline, while stops typically do. Consequently, a series [mamama] would often be categorized by the algorithm as a single SVI, while [papapa] would be categorized as three SVIs within a single SCU.

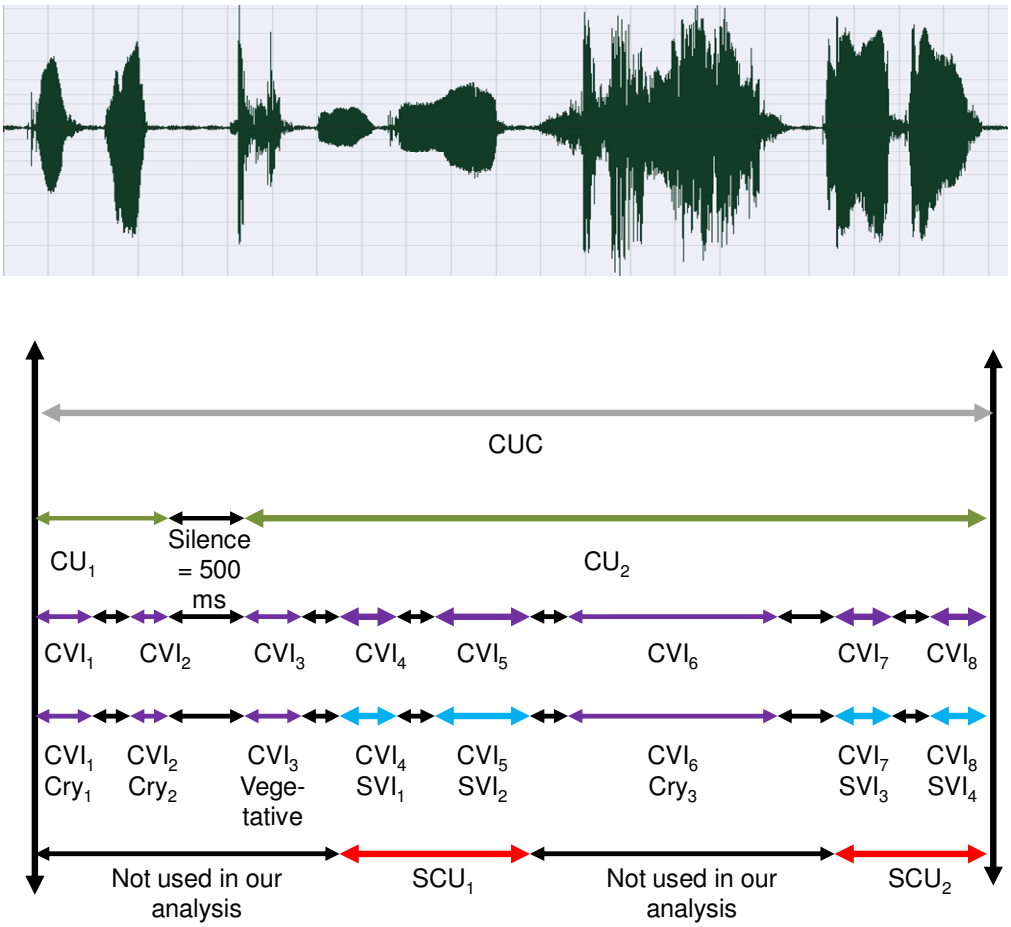

Figure S7. An illustration that a single CU can be broken up into more than one SCU by intervening sounds such as cries (produced by the key child), which are not deemed to be speech-related. In the pictured CUC, two CUs are separated by 500 ms of silence. The two CVIs in CU_1 are cries, and so are not categorized as corresponding to an SCU. Within CU_2, however, there are two SCUs, the first consisting of SVI_1 and SVI_2, and the second consisting of SVI_3 and SVI_4. These two SCUs are separated by Cry_3. Cries and vegetative sounds are not addressed in the acoustic analysis that is the primary focus of this work. See Supporting text for details of the procedure of identification of utterances and SVIs at all five steps.

The number of syllables in utterances as identified by a human listener was considerably higher than that identified by the machine algorithm based on the VC measure, because the algorithm's criterion for low energy boundaries between islands was set in such a way that some consonants (nasals and glides, for example) were often treated as within-island acoustic events rather than as indicators of island boundaries. Thus the concept "island" as we use it here has only a rough relation with the notion "syllable". The current algorithm for island identification has the advantage of being computationally efficient, requiring little CPU time. In future automated applications we intend to use models that are currently much less efficient computationally, but that more directly and completely assess syllabification.

The importance of SCU grouping owes to the fact that the acoustic analysis ultimately focused on features occurring in SVIs as a proportion of SCUs. The ratio measure (SVIs/SCUs) normalized the data on acoustic feature usage across children for differences in volubility or talkativeness, and because cries and vegetative sounds were not included in SCUs, the normalization ignored differences between the children in terms of likelihood to cry or produce other sounds unrelated to speech such as coughing or sneezing.

***Step 6: Automated acoustic feature analysis*.** In the sixth step the SVIs were analyzed acoustically: A 1024 point Fast Fourier Transform (FFT) was performed for each 64 ms frame in the analyzed recordings, and was advanced by 10 ms, whereupon the procedure was repeated until the entire SVI had been analyzed. At each frame, maximum spectral energy and total energy were determined in dB along with frequency at which the maximum spectral energy occurred. In addition entropy of the FFT spectrum was computed, and total energies were determined for 8 frequency bands of 1 kHz each. Also a 7-coefficient Linear Predictive Coding (LPC) spectrum was computed for each 15 ms frame and was advanced in 10 ms increments during analysis. F1, F2 and F3 frequencies were projected by the analysis to occur in ranges appropriate for a very young child vocal tract length (nominally 7 cm). An autocorrelation-based algorithm also determined pitch within the range of 60 to 1600 Hz at each 15 ms frame.

The final analysis was conducted at the level of the SVI. In order to be included in that analysis, each SVI had to fall within the durational range of 110 to 3000 ms. SVIs outside this range were not included in the analysis on the grounds that they were either too short to constitute well-formed syllables or too-long to constitute well-formed individual prosodic phrases in speech (64). The durational constraints were thus designed to confine the analysis to SVIs that could be expected to have greatest relevance to the development of speech.

From this point the automated analysis focused on classifying SVIs in the specified duration range dichotomously (positive classification or negative classification) with regard to the set of 12 acoustic parameters that had been selected to represent dimensions of infraphonological development (i.e., development of the infrastructure for speech) in infancy and early childhood. Table S5 outlines the 12 parameters.

The 12 acoustic parameters

1. The first group of parameters (the rhythmic/syllabification group) identified voicing events, canonical syllables, and moderately high spectral entropy, typical of speech (VC, CS and SE).
    a. VC or voicing (i.e., voiced SVIs per SCU) can be portrayed as a measure of the degree to which SCUs were acoustically organized to yield clear boundaries between periods of high energy phonatory regularity (voicing) and periods lacking that high energy voicing. In practice the measure roughly captured a minimum number of voiced islands (very roughly syllables) per SCU. Voicing was determined by whether the autocorrelation method was able to assign a pitch during each frame. If at least 10 frames or 60% of all frames in an SVI were assigned a pitch with 90% confidence, the SVI was categorized as positive for VC, otherwise it was categorized as negative.
    b. CS or canonical syllables per sequence provided a measure of the well-formedness of the initial formant transitions of each SVI with respect to initial transitions of syllables in mature speech (71). For SVIs 110 to 600 ms in duration, first and second formant frequencies were automatically tracked based on the LPC values from the beginning of each SVI until a maximum slope change for each formant could be determined. If, to that point, F1 slope was equal to or greater than 3 or F2 slope was equal to or greater than ± 5, if the max slope change had been reached within 120 ms, if the island had been designated as VC with average pitch > 250 and < 600 Hz, and if the bandwidths of F1 and F2 were lower than an empirically determined maximum, then the island was classified as positive for containing an initial CS. These criteria represent an approximation to the traditional acoustic specifications for canonical syllables in the infant vocalization literature (72).
    c. The classification SE was applied to islands with spectral entropy of the FFT exceeding an empirically determined value representing a deviation from the pattern of variation that is associated with pure normal phonation in childhood. The threshold for classification was set low enough that islands showing the sort of spectral variability in entropy that occurs in utterances of typical speech were classified as SE.
2. The second group of parameters (the low spectral tilt and high pitch control group) was represented by the number of islands per sequence showing squeal quality (a technical term in the infant vocalization literature), low spectral tilt, or a first spectral peak at high frequency (SQ, LT, HF):
    a. SQ represented pitch substantially exceeding a maximum value expected for child voices in speech-like utterances, set nominally at a mean of 600 Hz for the island.

b. LT was evidenced by high energy in the highest spectral band (7-8 kHz) relative to the energy at the maximum spectral peak from 0-6 kHz. If the highest band's energy was within 30 dB of the maximum spectral peak from 0 through 6 kHz for 50% of the island's frames, the island was classified as LT.
c. HF required the first (lowest frequency) spectral peak to occur at above 1.5 kHz for 25% of the island's frames.

3. The third group (the wide formant bandwidth and low pitch control group) was represented by the number of islands per sequence showing growl quality (again a technical term in the infant vocalization literature) or high bandwidth of resonances (GW, WB):
    a. GW required pitch to be substantially below an expected minimum level for infant/child voices in speech-like utterances, viz., mean < 250 Hz for the island.
    b. WB required that bandwidths for the first two formants (determined by a 3 dB drop from peak amplitude) exceed a value empirically determined to correspond to typical bandwidths of vowel-like sounds produced with normal phonation by children (> 400 Hz for F1and > 600 Hz for F2).
4. The fourth group (the duration group) was represented by the number of islands per sequence showing short, medium, long or extra long durations (S, M, L, XL):
    a. S islands were greater than 110 through 250 ms
    b. M islands were greater than 250 through 600 ms
    c. L islands were greater than 600 through 900 ms
    d. XL islands were greater than 900 through 3000 ms

The infrastructural parameters were derived from a theory (72) intended to be language universal. Infants all over the world are seen within this theory as beginning life with similar inclinations to explore the vocal capacity, and with similar anatomical and physiological capabilities with which to engage in that exploration. The parameters provide a frame for evaluation of the extent to which infant vocalizations produced in this exploration and in social interaction come to approximate infrastructural characteristics of well-formed, mature speech syllables (73). Since the most commonly occurring syllables in languages all over the world are drawn from a relatively small repertoire, all formed in accord with restrictions regarding the parameters above (voice, canonicity of formant transitions, spectral entropy and so on), the theory is able to provide a framework within which vocal development can be tracked universally early in life. Research within the model has addressed a variety of ambient languages (68, 70, 97, 72) (see below, Supporting Background: **Research in development of vocal acoustic characteristics).**

a. Rhythm/Syllabicity			Positive classification on group **a** features suggested speech-like rhythmic organization because values analyzed were islands (roughly, syllables) per utterance (SVIs per SCU) showing group **a** features. Thus, utterances were rhythmically organized in accord with speech if they tended to show relatively high numbers of syllables per utterance with voicing, canonical formant transitions, and spectral entropy variations typical of speech.
1	VC	Voiced or unvoiced: pitch detectable through > 50% of SVI (roughly, syllable)	
2	CS	Canonical Syllable transitions or not: Formant (F1 and F2) transitions < 120 ms	
3	SE	Spectral Entropy typical of speech or not	
b. Low spectral tilt and high pitch control			Positive classification on group **b** parameters suggested control of high pitch and low spectral tilt, which tend to occur in certain typical emotional expressions of high intensity (squeal quality). More islands per utterance positive on **b** parameters suggested more active emotional expression in the high spectral frequency range.
4	SQ	Mean pitch high or not (SQueal): > 600 Hz	
5	LT	Low Tilt of spectrum or not	
6	HF	High Frequency energy concentration or not	
c. Wide formant bandwidth and low pitch control			Positive classification on group **c** parameters suggested control of low pitch and high bandwidths of the first two formants, qualities which tend to occur in certain typical emotional expressions of high intensity (growl quality). More islands per utterance positive on **c** parameters suggested more active emotional expression in the low spectral frequency range.
7	GW	Mean pitch low or not (GroWl): < 250 Hz	
8	WB	Wide Bandwidth of first two formants or not	
d. Duration of islands (SVIs) within utterances (SCUs)			Group **d** parameters split according to durations typical of syllables in speech. Positive classification on parameters 9 and 10 suggested speech-like rhythmic organization because the durational values indicated are typical of syllables in speech. More islands per utterance with 9 and 10 thus suggested more speech-like syllables. Positive classification on parameters 11 and 12 suggested the opposite, because the corresponding ranges are beyond the durations of typical syllables.
9	S	Short (110 - 250 ms)	
10	M	Medium (250 - 600 ms)	
11	L	Long (600 - 900 ms)	
12	XL	EXtra Long (900 - 3000 ms)	

Table S5: The 12 acoustic parameters used for automated analysis of SVIs. These pertain to four groupings a-d, indicated also by color coding. See accompanying text above for further explication of the parameters.

Reliability of the automated analysis

Reliability of identification of the key child voice. To assess the reliability of the automated analysis, a 70-hour sample from the recordings was coded auditorily by a panel of phonetically trained listeners and used to test the automated labeling. One-hour samples were selected from 70 different infants/children distributed across the age range from 2-36 mo. A reliability analysis related to this one can be found in LENA Foundation technical reports and has been reported by Xu et al. (45). The present analysis is based on a slightly different focus: Both "near" (high signal to noise ratio) and "far" (low signal to noise ratio) CUCs were included in the reliability analysis here, whereas in Xu et al., only near CUCs were included. The 70 hours had been selected from recordings at 2-36 months of age, with each hour composed of six different non-contiguous periods of high vocal activity. The reliability analysis was based on comparisons of 10 ms time periods across the human transcribed and machine labeling, implemented with a 30 ms. collar guard, also specific to the present study.

Periods that were automatically labeled in step one above as pertaining to a CUC were found to agree with the human transcriber labeling 73% of the time, with only 5% false positives (see Table S6a). Similarly, 64% of the time periods identified by the algorithm as pertaining to CUC, were also identified by the human listeners as pertaining to CUCs (see Table S6b). Because a larger proportion of all intervals were "other" than were "key child", the absolute numbers of "other" errors were higher than might be expected from the percentages. In fact false positives in Table S6a, while representing only 5% of identified "others", occurred in absolute numbers of intervals about 1/3 as often as true positives (hits).

a.

Human listener classification	Machine classification	
	Key child	Other
Key child	0.73	0.27
Other	0.05	0.95

b.

Human listener classification	Machine classification	
	Key child	Other
Key child	0.64	0.03
Other	0.36	0.97

Table S6. Proportion correct classification of all 10 ms frames as belonging to CUCs (key child utterance clusters) as opposed to belonging to other sounds or silences based on 70 hours of auditorily classified (by human transcribers) and machine classified data with a 30 ms. collar guard. In these data, the CUCs (Key child) were strictly limited to utterances labeled as pertaining to the child wearing the recorder. Segments labeled as "overlapping" with other sounds or voices by the automated labeling were, of course, not included. **(a)** Classification results based on computation of proportions using row sums of raw numbers of 10 ms frames as denominators (gold standard = human listener). **(b)** Classification results based on column sums (gold standard = machine). See Supporting text for further details on reliability analysis procedures and results.

Given the state of the present technology, human listeners prove to be better than the machine algorithm at recognizing voices when there is overlap or background noise, a fact that accounts for the relatively high false negative rate of the machine algorithm (Table S1a). The algorithm is unable to be as certain as human listeners that an unusual sound or a sound in noise might be the product of the child voice. The goal of the automated identification procedure is to maintain high certainty on vocalizations that are identified as being produced by the child, and thus an important feature of the automated maximum likelihood procedure is its relatively low false positive rate, ensuring that when it identified a child vocalization, it was predominantly correct.

In a separate reliability test, the 70 hours of auditorily coded material were used to test the automated labeling of CVIs as SVI vs other (cry/laugh, vegetative). This analysis focused on CVIs that had been identified both by the machine algorithm and the human transcribers. The hit rate of the machine algorithm for SVIs identified by human listeners averaged 75% across the age range (see Table S7a), indicating that SVIs were generally well-differentiated from the infant's own cry and vegetative sounds: Confusions were primarily associated with fussy sounds, the kinds of sounds that include characteristics of both speech-like sounds and cry. Similarly 86% of SVIs as indicated by the machine algorithm were also identified as SVIs by the human listeners (see Table S7b).

a.

Human listener classification	Machine classification	
	SVI	Cry/Vegetative
SVI	0.75	0.25
Cry/Vegetative	0.16	0.84

b.

Human listener classification	Machine classification	
	SVI	Cry/Vegetative
SVI	0.86	0.28
Cry/Vegetative	0.14	0.72

Table S7. Proportion correct classification of all 10 ms frames within CVIs as either SVI or cry/vegetative for 70 hours of auditorily classified (by human transcribers) and machine classified data. (a) Classification results based on computation of proportions using row sums of raw numbers of frames as denominators (gold standard = human listener). **(b)** Classification results based on column sums (gold standard = machine). See Supporting text for further details on reliability analysis procedures and results.

Additional reliability checks were conducted with a sample of data from all three groups of infants/children evaluated in the present work (typically developing, autistic, language delayed) across the range from 9 to 41 months. 16 five-minute periods were selected, one each from 8 infants and children in the typically developing sample, and 4 each from infants and children in the autistic and language delayed samples. The listener/human transcriber was the first author, who has been involved in research on infant vocal

development for over 30 years (73, 74). This research has involved the development of a widely utilized categorization scheme for early infant sounds (72) as well as extensive research in phonetic transcription and acoustic analysis of infant and child speech (75, 76). He has trained hundreds of students and colleagues in categorization of speech and speech-related vocalizations and in phonetic transcription, and has served as the "gold-standard" transcriber/coder in published research in infant vocalizations (77).

One human/machine agreement evaluation based on the 16 samples focused on the automatically identified SVIs. The listener made auditory judgments on 1202 automatically identified SVIs occurring in these 16 samples. The first judgment in each case was whether the SVI had in fact been produced by the infant/child or by some other sound source. Fewer than 0.03 of the SVIs failed that auditory test, suggesting again a very low false positive rate in the automated identifications – vocalizations identified by the automated procedure as produced by the child were indeed overwhelmingly produced by the key child, although there were cases where the key child's voice was not the only one that could be heard during the SVI interval. The low false positive rate is a further indicator of the conservatism of the automated approach with regard to infant/child voice identification.

Reliability of automated acoustic feature identification. The same 1202 SVIs from the three groups of infants/children were evaluated by auditory and acoustic inspection individually. Duration measurements yielding the fourfold classification (S, M, L, and XL) were correct in portraying labeled intervals, but it should be noted that labeled SVI intervals often represented less than the total continuous vocalization period of infant/child utterance as heard at playback. In practice, machine-identified SVIs often represented beginning and middle portions of auditorily identified infant vocalizations, where amplitude was high, phonation was relatively normal and low in noise, and where overlapping sounds were minimal to nil.

Additional analysis of the 1202 vocalizations was used to evaluate the automated results against auditory judgments made by the human transcriber on five of the acoustic parameters (VC, CS, SE, SQ, GW). In each case the observer located vocalizations by referencing the labels provided by the algorithm, ignored the vocalizations whose labels indicated they were cries or vegetative sounds, had been produced by another speaker, had failed to meet the duration criteria, or were not SVIs, and then coded the remaining SVIs in terms of the five designated parameters. The algorithm precluded the possibility that an SVI coded as either SQ or GW could also be coded as CS. Also an SVI could not be coded as both SQ and GW.

In each case, prior research has addressed characteristics of infant/child vocalizations that are closely related to these five parameters. In fact they had been selected and designed as initial attempts to implement acoustic tracking for key characteristics of infant/child vocalization. The goal of this aspect of the reliability evaluation, then, was to assess the degree to which the automated acoustic feature classifications were reliably related to intuitively-based auditory classifications that are in common use in research on infant vocal

development. The other features (LT, HF and WB) were also selected as being related to features of child vocalizations, but they have not been the focus of prior auditory judgments by human transcribers to our knowledge, and were not subjected to such judgments here.

VC reliability (for the data see Table S*8).* For VC the human listener judgments were based on whether each labeled SVI interval was predominantly voiced. Because such a large proportion of SVIs were VCs, the measure of VC per speech-related child utterance (SCU) can be thought of as a reliable indicator of the degree of within-SCU organization, a very rough indicator of the degree of syllabic-like organization within key child breath groups.

Acoustic parameter	Session level correlation between human listener and machine classifications	Proportion correct machine classifications with human listener as gold standard	Cohen's kappa for human listener vs. machine classification	Chi-square probability of Cohen's kappa
VC	0.99	0.99	0.78	<.0001
CS	0.52	0.65	0.21	<.0001
SE	0.87	0.82	0.64	<.0001
SQ	0.46	0.91	0.25	<.0001
GW	0.72	0.9	0.48	<.0001

Table S8. Reliability results for 5 of the 12 acoustic parameters (those that could be judged auditorily in accord with procedures of prior research in infant/child vocalization)**.** In column one, session-level agreement is recorded across the 16 five-minute samples (8 typically developing, 4 autistic, 4 language-delayed), represented by the correlation between the number of positive judgments made for each of the 16 samples by the human listener (transcriber) for each parameter and the number of positive judgments on that parameter made by the machine algorithm for the same 16 samples. The remaining columns pertain to two-by-two classification tables (for example, VC vs. not VC, for human listener vs. machine classification, where human listener is taken as the gold standard). Proportions correct are high but misleading because of cell imbalances. Cohen's kappa corrects for these imbalances of distributions. The chi-square probability provides a measure of the significance of agreement between the auditory and automated classifications whenever kappa is positive (78). Even relatively low levels of agreement (as indicated by kappa values) between human listener and machine recognition appear to be capable of providing important measures of the acoustic organization of infant and child vocalizations, as indicated in our results, perhaps due to the huge sample size we analyzed compared with sample sizes in prior research on vocal development. See Supporting text for further details on reliability analysis procedures and results.

CS reliability (see Table S8). Both auditory and machine judgments categorized only the first detected syllable of each SVI as CS or not CS. The auditory procedure was similar to that employed routinely in identification of canonical babbling in infancy (79, 80), with assignment of the label to all canonical syllables of CV or CCV shape regardless of the nature of perceived consonant-like or vowel-like components of the syllables. The machine algorithm's hits on CS (true positive CS) were 1.3 times more frequent than false positives, but false negatives were 1.6 times more frequent than hits, indicating that the machine algorithm tended to miss many auditorily perceived instances of canonical syllables. Since the measure analyzed in the results of our study is the ratio of CS judgments (for SVIs) to SCUs, the value represents a rough indicator of (a lower bound on) the number of canonical syllables per breath group.

SE reliability (see Table S*8).* The SE auditory judgments were made on the basis of perceived voice quality. For each of the 16 samples, the judge first listened to enough utterances to gauge each infant or child's normal voice quality at their habitual pitch level (81), in the middle of stressed syllable nuclei. SVIs with substantial perceived roughness, harshness or noisiness (including substantial breathiness) (82), were judged positive on SE, while SVIs that were dominated by normal voice relatively free of noise were judged negatively. It is important to emphasize that the judgments were made with a "low threshold": The intent was to include in the SE category SVIs that showed spectral aperiodicity of a sort that would be expected to occur regularly in the utterances (breath groups) of normal speech. The low threshold yielded many more judgments of SE based on rough or harsh voice than of GW based on rough or harsh voice, where the criterion threshold of roughness was decidedly higher. The algorithm performed much better than chance in terms of agreement with the human listener, and yielded a measure (ratio of SVIs with SE to SCUs) of the degree to which utterance-level organization showed normal vocal quality variability.

SQ reliability (see Table S*8).* For SQ the machine algorithm targeted SVIs with average pitch of 600 Hz or higher. The auditory "squeal" judgments, on the other hand, were based on the common laboratory listening method which relies on an intuitive pitch criterion. In accord with the criterion, SVIs that were perceived as having a salient period that was above the habitual pitch range of the infant/child were categorized by the listener as SQ. The auditory criterion yielded a 3.6 times as many SQ judgments as the machine algorithm. Still, hits outnumbered false positives in a ratio of 1.8.

GW reliability (see Table S*8).* The machine algorithm categorized SVIs with low pitch (average 250 Hz or lower) as GW. As with SQ, the listener judgments were based on the common laboratory criterion: Any SVI that had a salient pitch characteristic lower than the individual's habitual pitch and also any SVI that had especially harsh or rough voice quality with pitch in the mid or low range was categorized as GW. Subharmonic energies appear to be particularly effective in inspiring listeners to apply a growl label to infant sounds (83). Note that the auditory criterion for harsh voice quality was higher for GW than SE. The judgment of GW was

made based on harshness if and only if the utterance was perceived as being out of the range of normal voice quality harshness in conversational speech.

SUPPORTING RESULTS

Mean values across the 12 parameters for each of the three child groups

Figure S8 supplies the mean raw values for presence of each parameter in the recordings, that is, the mean number of occurrences of SVIs designated as "plus" for each of the 12 parameters for each recording. Here there is no normalization for volubility or length of recording. Because of the lack of normalization, these values were not utilized in comparing groups statistically, and are provided only to illustrate raw amounts of usage for the various parameters.

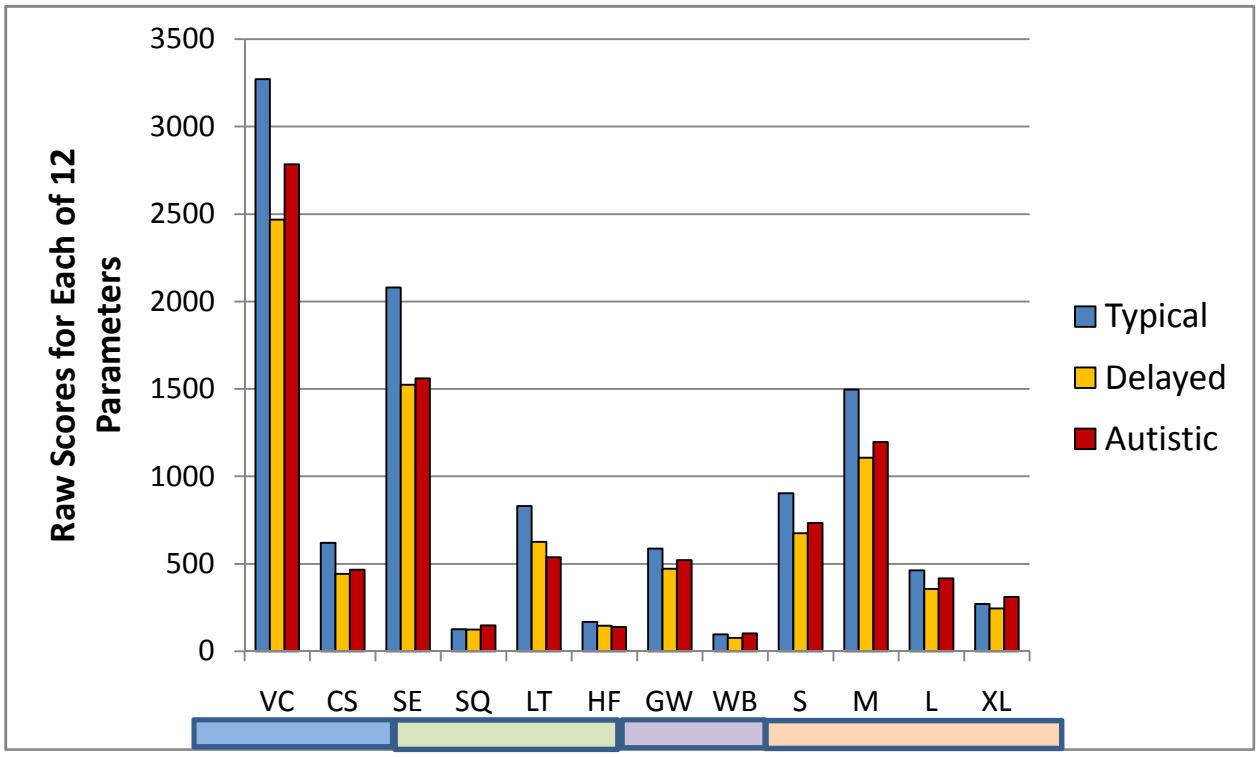

Figure S8. The average number of SVIs classified as "plus" for each of the 12 acoustic parameters for each recording (typically developing sample N = 802, autism sample N = 351, language delayed sample N = 333). Note the wide range of values. Also note similarities in the pattern of values for the 12 parameters across groups. The normalized values in Figure S9 adjust for differences across children in volubility (rate of vocalization) and for differences in length of recordings.

In Figure S9, the data are provided on the ratio scores (SVI/SCU) for each parameter. These values suggest notable group differences. ANOVA with Posthoc Tukey's t comparisons illustrating that indeed the groups differed robustly on several parameters, and especially that the typically developing sample tended to differ from the autism sample.

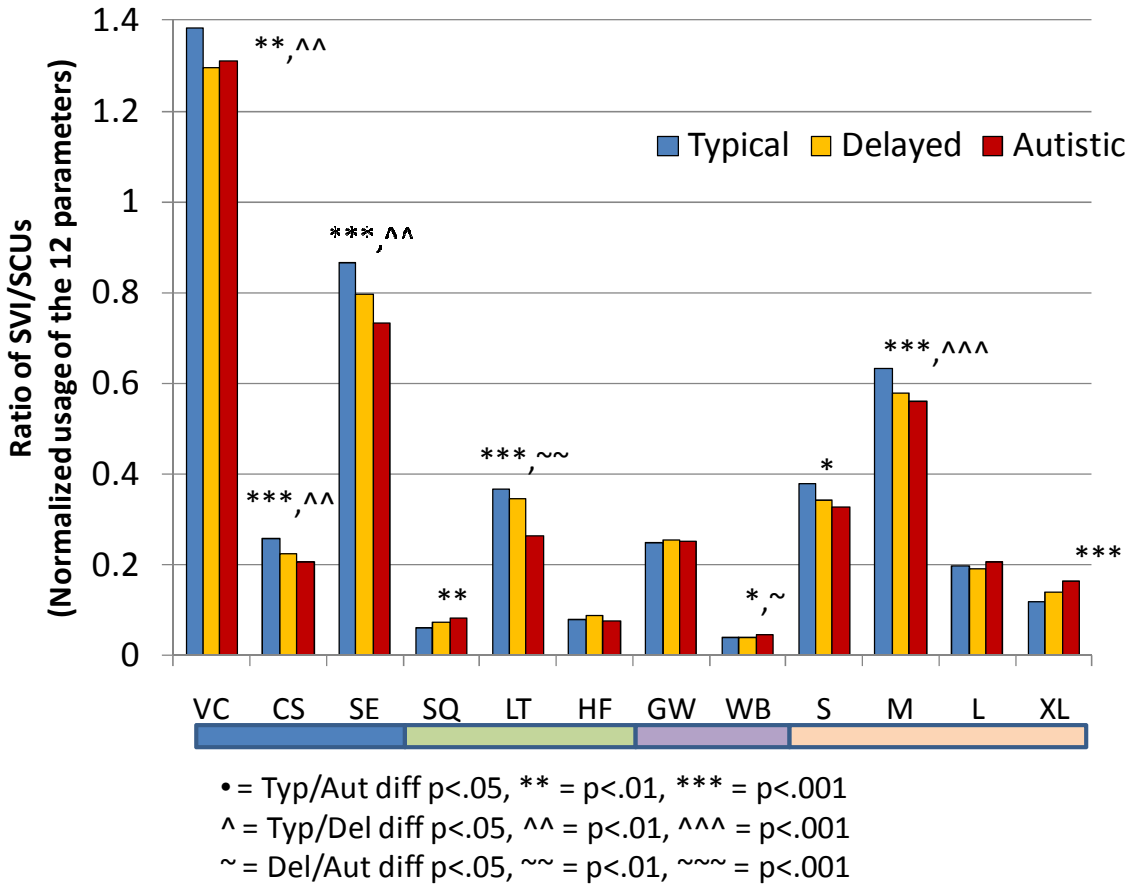

Figure S9: Mean usage by child group of the 12 acoustic parameters as determined by the automated analysis, normalized for volubility and recording length by representing usage in terms of the ratio of SVIs to SCUs. Typically developing and autism samples differed statistically reliably on SVI/SCU ratios for 9 of the 12 parameters, providing an initial indication of the potential of the automated acoustic analysis for group differentiation. Typically developing and language delayed samples differed reliably on four parameters.

Figure S10 provides further evidence of group discriminability. Here the scores for each of the 12 parameters were normalized within one-month age groupings, correcting for the fact that the samples were not distributed evenly across ages for the three groups. See Figure 1 in the main text or Table S3 for age differences across groups. The age normalization is an important step prior to discriminant analysis (see above **Statistical analysis summary**).

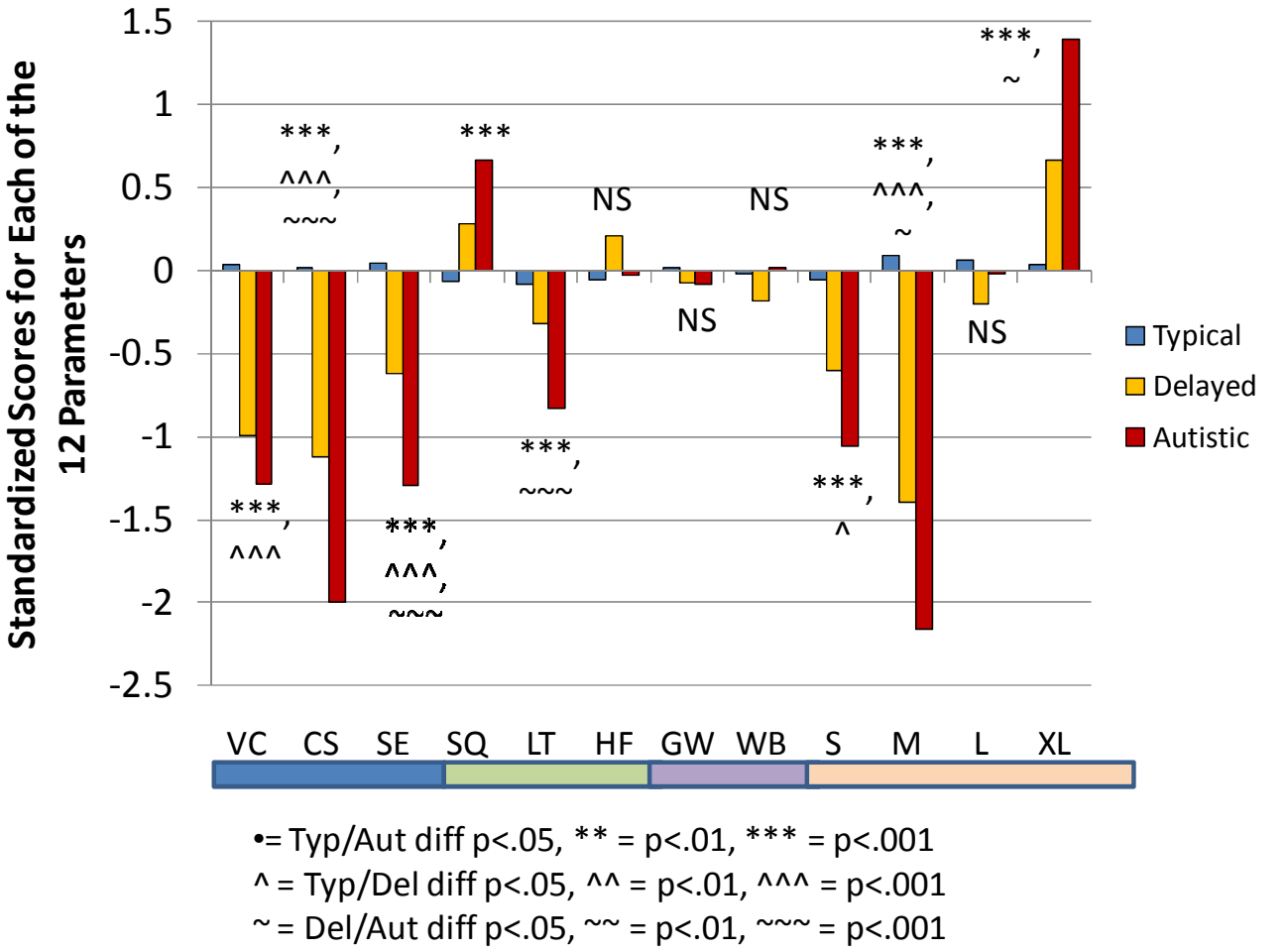

Figure S10. Group comparisons based on means for each child group on the standardized scores (i.e., z-scores) for the 12 acoustic parameters. The z-scores were computed for the typically developing group at the recording level by first determining means and SDs at each month of age (windows four-months wide with two months of overlap). These means and SDs for the typically developing sample were used to compute z-scores for all three groups within each of the age ranges at the recording level. Then mean z-scores were computed at the child level (across recordings for each child).These z-scores were entered into LDA for modeling group discrimination. The results in Figure S10 illustrate that the most prominent differences among groups were on parameters of the rhythmic/syllabicity grouping (VC CS and SE) along with the duration parameters (except L). Significance levels are indicated in the same ways as in Figure S9. The z-score means for the typically developing group in Figure S10 are always near zero because the typically developing group was the reference group with a mean of 0 and SD of 1 in the computational procedure. The z-score means are not exactly zero for the typically developing group because the z-scores were first established at the recording level, and then averaged across child (with children differing in numbers of recordings).

Correlational results indicating empirical and theoretical organization of the parameters

The 12 parameters were grouped for the analyses *a priori* in part in accord with theoretical considerations and research in infant vocalizations that have been discussed above. However, the grouping was also developed to reflect the results of correlations (and see below, **Principal components analysis indicating**

empirical and theoretical organization of the parameters) conducted on the recorded data to provide an empirical assessment of the associations among the parameters and to help support interpretation of the results. Tables 9a-f provide the correlational results and levels of significance for each parameter, with panels broken down by child group. In addition the tables show the correlation of each of the 12 parameters with age of children.

a. Typically developing sample, correlations													
	Age	VC	CS	SE	SQ	LT	HF	GW	WB	S	M	L	XL
VC	***0.68***												
CS	***0.63***	**0.79**											
SE	***0.51***	**0.65**	**0.62**										
SQ	-0.20	-0.23	-0.29	-0.15									
LT	0.11	0.14	0.18	**0.50**	0.08								
HF	-0.14	-0.23	-0.17	0.15	**0.50**	**0.49**							
GW	0.21	**0.45**	0.18	0.37	-0.16	0.00	-0.21						
WB	***0.46***	**0.63**	**0.51**	**0.44**	-0.22	0.12	-0.25	**0.67**					
S	***0.44***	**0.78**	**0.74**	**0.52**	-0.02	0.21	-0.10	**0.51**	**0.66**				
M	***0.68***	**0.75**	**0.76**	**0.61**	-0.30	0.13	-0.15	0.20	0.36	**0.42**			
L	0.25	0.24	-0.10	0.09	-0.21	-0.04	-0.11	-0.08	-0.06	-0.33	0.20		
XL	-0.20	-0.15	**-0.49**	-0.24	-0.07	-0.17	-0.08	-0.15	-0.23	**-0.52**	-0.36	**0.62**	
b. Typically developing group, statistical significance levels													
VC	***0.0000***												
CS	***0.0000***	**0.0000**											
SE	***0.0000***	**0.0000**	**0.0000**										
SQ	0.0000	0.0000	0.0000	0.0000									
LT	0.0025	0.0001	0.0000	**0.0000**	0.0178								
HF	0.0001	0.0000	0.0000	0.0000	**0.0000**	**0.0000**							
GW	0.0000	**0.0000**	0.0000	0.0000	0.0000	0.9007	0.0000						
WB	***0.0000***	**0.0000**	**0.0000**	**0.0000**	0.0000	0.0009	0.0000	**0.0000**					
S	***0.0000***	**0.0000**	**0.0000**	**0.0000**	0.6598	0.0000	0.0062	**0.0000**	**0.0000**				
M	***0.0000***	**0.0000**	**0.0000**	**0.0000**	0.0000	0.0002	0.0000	0.0000	0.0000	**0.0000**			
L	0.0000	0.0000	0.0031	0.0114	0.0000	0.3004	0.0018	0.0178	0.0702	0.0000	0.0000		
XL	0.0000	0.0000	**0.0000**	0.0000	0.0354	0.0000	0.0284	0.0000	0.0000	**0.0000**	0.0000	**0.0000**	

Table S9a-b. (a) Correlations for typically developing sample and (b) statistical significance levels. Color coding for groupings of the acoustic parameters correspond to Figure 2a in the main text and Table S5: **blue = r**hythmic/syllabification (RhSy)**, green =** spectral tilt/high pitch (LtHp)**, violet =** bandwidth/low pitch (BwLp)**, peach =** duration (Dur). Notice 6 correlations of acoustic parameters with age are > 0.4, and all correlations with age are statistically significant. **Boldface** indicates correlations > ±0.4. Correlations between pairs of the 12 acoustic parameters (that is, correlations not involving age) showed notable similarities for all three child groups, and these similarities tended to support the theoretical groupings of the parameters. The correlations among members of each of the four *a priori* parameter groupings are displayed in the tables with

background colors corresponding to the groupings. Note that all correlations within the blue parameter grouping, RhSy (VC, CS, SE), are > 0.6, and several within the other color-coded groupings are also high. All three child groups showed correlations > ±0.4 (a) for all three pairings of parameters of the rhythmic/syllabicity grouping (VC with SE, CS with SE, VC with SE), (b) for two pairings of the spectral tilt/high pitch grouping (SQ with HF, HF with LT), (c) for the single pairing of the bandwidth/low pitch grouping (GW with WB), and (d) for at least 3 of the 6 pairings of the duration grouping.

c. Language delayed sample, correlations													
	Age	VC	CS	SE	SQ	LT	HF	GW	WB	S	M	L	XL
VC	***0.52***												
CS	***0.56***	**0.75**											
SE	0.37	**0.64**	**0.63**										
SQ	-0.01	0.11	-0.15	0.16									
LT	0.15	0.26	0.27	**0.65**	0.16								
HF	-0.10	-0.04	-0.03	0.36	**0.49**	**0.52**							
GW	0.11	0.33	0.13	0.34	-0.03	0.18	-0.03						
WB	0.28	0.39	0.35	0.28	-0.16	0.16	-0.13	**0.73**					
S	***0.46***	**0.81**	**0.74**	**0.56**	0.17	0.31	0.02	**0.48**	**0.53**				
M	***0.53***	**0.64**	**0.79**	**0.61**	-0.10	0.21	0.03	0.14	0.26	**0.43**			
L	-0.15	-0.13	**-0.42**	-0.19	-0.05	-0.13	-0.06	**-0.43**	**-0.41**	**-0.60**	-0.22		
XL	***-0.49***	**-0.45**	**-0.78**	**-0.52**	0.01	-0.26	-0.05	-0.27	**-0.40**	**-0.65**	**-0.75**	**0.65**	
d. Language delayed sample, statistical significance levels													
	Age	VC	CS	SE	SQ	LT	HF	GW	WB	S	M	L	XL
VC	***0.0000***												
CS	***0.0000***	**0.0000**											
SE	0.0000	**0.0000**	**0.0000**										
SQ	0.8391	0.0541	0.0066	0.0035									
LT	0.0073	0.0000	0.0000	**0.0000**	0.0043								
HF	0.0588	0.4221	0.5508	0.0000	**0.0000**	**0.0000**							
GW	0.0431	0.0000	0.0203	0.0000	0.6488	0.0010	0.5718						
WB	0.0000	0.0000	0.0000	0.0000	0.0031	0.0038	0.0154	**0.0000**					
S	***0.0000***	**0.0000**	**0.0000**	**0.0000**	0.0021	0.0000	0.7179	**0.0000**	**0.0000**				
M	***0.0000***	**0.0000**	**0.0000**	**0.0000**	0.0747	0.0001	0.5790	0.0099	0.0000	**0.0000**			
L	0.0064	0.0142	**0.0000**	0.0004	0.3252	0.0157	0.2555	**0.0000**	**0.0000**	**0.0000**	0.0000		
XL	***0.0000***	**0.0000**	**0.0000**	**0.0000**	0.8465	0.0000	0.3653	0.0000	**0.0000**	**0.0000**	**0.0000**	**0.0000**	

Table S9c-d. (c) Correlations for language delayed sample and (b) statistical significance levels. Color coding for groupings of the acoustic parameters correspond to Figure 2a in the main text and Table S5: **blue =** rhythmic/syllabification (RhSy)**, green =** spectral tilt/high pitch (LtHp)**, violet =** bandwidth/low pitch (BwLp)**, peach =** duration (Dur). **Boldface** indicates correlations > ±0.4. Notice 5 correlations of acoustic parameters with age are > ±0.4, and, as in the case of the typically developing sample, all correlations with age are statistically significant. Also note that, as in the typically developing sample, all correlations within the blue

parameter grouping, RhSy (VC, CS, SE), are > 0.6, and several within the other color-coded groupings are also high.

e. Autism sample, correlations													
C	Age	VC	CS	SE	SQ	LT	HF	GW	WB	S	M	L	XL
VC	0.13												
CS	0.09	**0.85**											
SE	0.08	**0.72**	**0.71**										
SQ	-0.09	-0.15	-0.29	0.13									
LT	0.19	0.21	0.21	**0.56**	0.15								
HF	-0.04	-0.05	-0.09	0.34	**0.50**	**0.57**							
GW	0.00	**0.53**	***0.41***	***0.60***	0.01	0.38	0.29						
WB	0.17	**0.64**	**0.62**	**0.50**	-0.11	0.29	0.07	**0.67**					
S	0.05	**0.84**	**0.83**	**0.71**	-0.07	0.32	0.10	**0.71**	**0.80**				
M	0.12	**0.70**	**0.77**	**0.72**	-0.12	0.24	0.01	0.31	0.29	**0.48**			
L	0.06	***-0.42***	***-0.63***	***-0.53***	0.04	-0.31	-0.14	***-0.61***	***-0.60***	**-0.77**	-0.33		
XL	-0.04	***-0.60***	**-0.80**	***-0.73***	0.01	-0.35	-0.15	***-0.56***	***-0.56***	**-0.75**	**-0.76**	**0.73**	

f. Autism sample, statistical significance levels													
	Age	VC	CS	SE	SQ	LT	HF	GW	WB	S	M	L	XL
VC	0.0167												
CS	0.0941	**0.0000**											
SE	0.1320	**0.0000**	**0.0000**										
SQ	0.0960	0.0048	0.0000	0.0150									
LT	0.0003	0.0001	0.0001	**0.0000**	0.0052								
HF	0.4514	0.3584	0.1021	0.0000	**0.0000**	**0.0000**							
GW	0.9453	**0.0000**	***0.0000***	***0.0000***	0.9210	0.0000	0.0000						
WB	0.0017	**0.0000**	**0.0000**	**0.0000**	0.0334	0.0000	0.1639	**0.0000**					
S	0.3466	**0.0000**	**0.0000**	**0.0000**	0.1670	0.0000	0.0676	**0.0000**	**0.0000**				
M	0.0192	**0.0000**	**0.0000**	**0.0000**	0.0204	0.0000	0.8728	0.0000	0.0000	**0.0000**			
L	0.2442	***0.0000***	***0.0000***	***0.0000***	0.4305	0.0000	0.0112	***0.0000***	***0.0000***	**0.0000**	0.0000		
XL	0.4512	***0.0000***	**0.0000**	***0.0000***	0.7926	0.0000	0.0060	***0.0000***	***0.0000***	**0.0000**	**0.0000**	**0.0000**	

Table S9e-f. (e) Correlations for autism sample and (f) statistical significance levels. Color coding for groupings of the acoustic parameters correspond to Figure 2a in the main text and Table S5: **blue =** rhythmic/syllabification (RhSy)**, green =** spectral tilt/high pitch (LtHp)**, violet =** bandwidth/low pitch (BwLp)**, peach =** duration (Dur). **Boldface** indicates correlations > ±0.4. As with the other child groups, all three correlations between parameters within the blue parameter grouping (VC, CS, SE) are > 0.6, and several within the other color-coded groupings are also high. However, in stark contrast to the other child groups, *no* correlations of acoustic parameters with age are > ±0.2 (first column), and statistical significance levels are much lower for the autism sample, indicating that children in the autism sample did not show age progression with the parameters to the extent the other groups did. Leaving aside the 12 correlations with age and the 13 correlations among parameters within the four theoretical groupings, there are 53 additional correlations displayed in each of the Tables S9a, S9c and S9e. Here the autistic sample showed 11 correlations above the

> ±0.4 criterion level (indicated in ***red bold italic font***), that were not above the criterion in the typically developing group and nine that were not above the criterion in the language delay group. We interpret these uniquely high correlations among acoustic parameters (outside the four *a priori* parameter groupings) in the autism sample (especially given their low correlations of the acoustic parameters with age) as indicators that the acoustics of vocalization is organized differently in the autism group than in the other groups.

MLR analysis on the acoustic parameter groupings

The same analysis that was performed for all 12 parameters in the main text Fig. 2b, was also conducted on the four *a priori* groupings of the 12 parameters, and these analyses are displayed in Figure S11a-d. See caption for explanation.

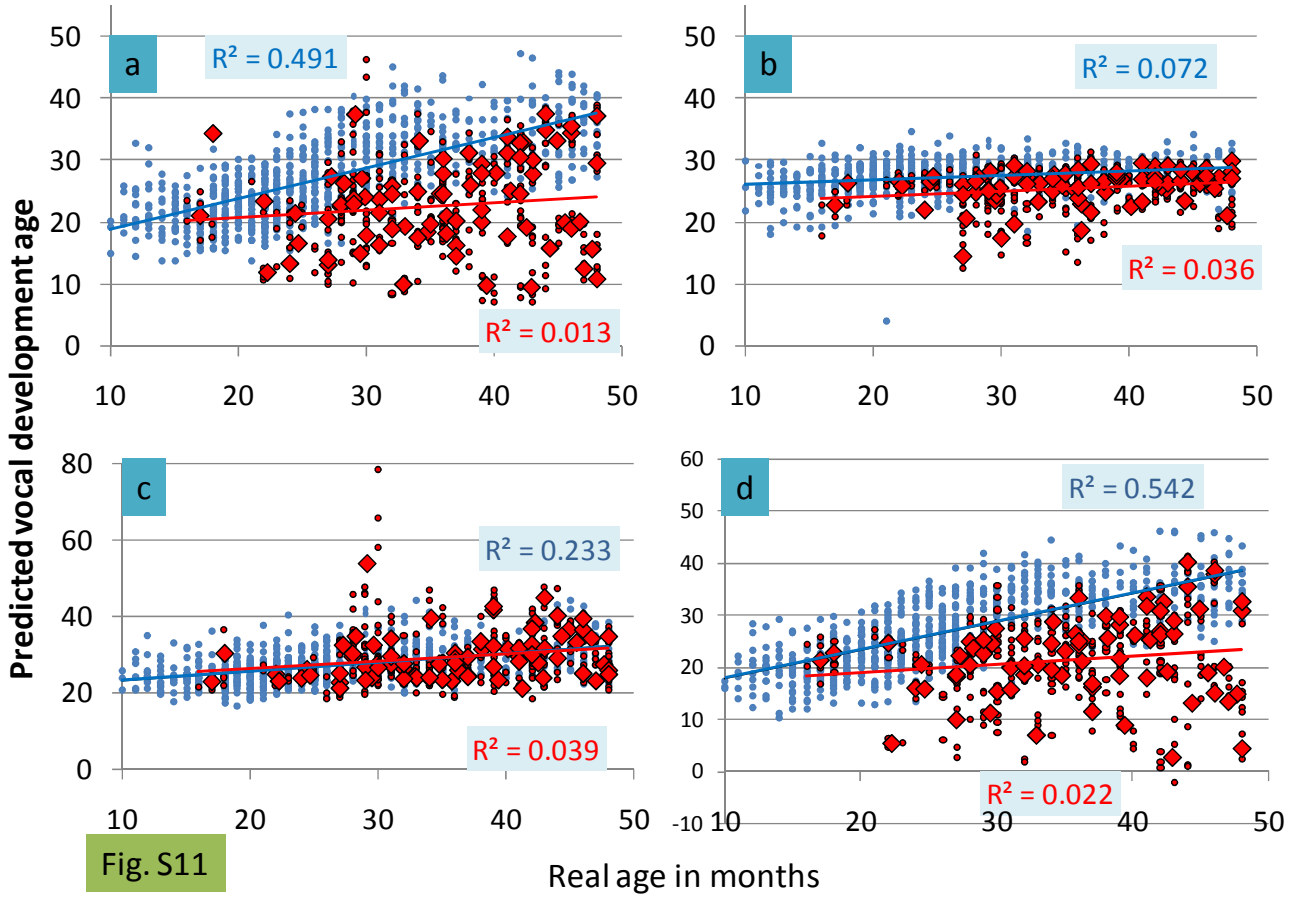

Figure S11: *MLR comparisons of the typically developing and autism samples for the four a priori groupings of 12 acoustic parameters*: **(a)** rhythm/syllabicity (RhSy); **(b)** low spectral tilt/high pitch (LtHp); **(c)** high bandwidth/low pitch (BwLp); **(d)** duration of SVIs (DUR). The MLRs show strong group differentiation in panels **(a),** **(b),** and **(d),** the autism sample having lower predicted ages than the typically developing sample, and in panels **(a)** and **(d),** the autism sample having lower correlations with age. The high bandwidth/low pitch grouping (panel **c**) shows little group differentiation.

Principal components analysis indicating empirical and theoretical organization of the parameters

Principal components analysis (PCA) for the three child groups in Table S10 and Figure S12a-c provided empirical support for the conceptual organization of the infrastructural acoustic parameter system (see above, **The 12 acoustic parameters**, and ***Step 6: Automated acoustic feature analysis***). The principal components that emerged, especially in the typically developing sample (see blue shaded correlations), were highly associated with the *a priori* conceptual groupings (see color coded left column, Acoustic parameters).

	Typically developing					Language delayed					Autism			
Acoustic parameters	PC1	PC2	PC3	PC4		PC1	PC2	PC3	PC4		PC1	PC2	PC3	PC4
VC	0.78	0.51	0.09	-0.05		0.81	0.32	0.04	0.05		0.49	0.74	-0.04	-0.10
CS	0.88	0.20	-0.30	-0.07		0.92	0.07	-0.08	-0.29		0.49	0.80	-0.04	-0.21
SE	0.71	0.36	0.07	0.39		0.75	0.28	0.46	0.10		0.39	0.74	0.41	0.15
SQ	-0.40	-0.03	-0.23	0.57		-0.12	-0.10	0.73	-0.20		-0.05	-0.06	0.14	0.96
LT	0.30	0.05	0.02	0.77		0.37	0.24	0.67	0.17		0.19	0.18	0.91	-0.02
HF	-0.12	-0.19	-0.06	0.87		0.01	-0.10	0.88	-0.02		0.09	-0.05	0.75	0.50
GW	0.05	0.91	-0.02	-0.08		0.07	0.92	0.04	-0.15		0.79	0.17	0.30	0.04
WB	0.34	0.82	-0.08	-0.09		0.23	0.84	-0.14	-0.19		0.85	0.22	0.07	-0.10
S	0.51	0.62	-0.44	0.07		0.62	0.47	0.14	-0.42		0.83	0.50	0.06	-0.01
M	0.91	0.07	-0.01	-0.08		0.88	-0.03	-0.05	-0.15		0.06	0.95	0.09	-0.06
L	0.20	-0.08	0.91	-0.06		-0.15	-0.32	-0.08	0.88		-0.79	-0.27	-0.11	-0.02
XL	-0.34	-0.05	0.85	-0.08		-0.69	-0.11	-0.03	0.62		-0.52	-0.71	-0.15	-0.06

Table S10. Results of principal components analysis (PCA) on the 12 acoustic parameters within the three child groups. Colored shading (with colors corresponding to groups) is used to indicate the parameters that correlated at > 0.55 with each component. Using this selection rule, *simple structure* can be observed in the typically developing sample, in that each principal component (PC) possessed a unique set of highly-correlated parameters, not shared with the other components. Further, the PCA organization in the case of the typically developing sample makes speech scientific sense, in that all the parameters conceptually associated with any of the first three *a priori* acoustic parameter groupings (indicated by the shading, light blue, green, violet) showed their highest correlations within that grouping. For the duration grouping (peach shading), the parameters split: High values on two of them (L and XL, i. e., long and extra long) characterized a separate PC (which might be termed "longer duration of SVIs than is normally found for syllables in mature languages"), while the other two parameters (S and M, i. e., short and medium duration) were associated with other PCs. The association of M with the rhythmic/syllabicity PC is predictable again on speech scientific grounds because M was defined by duration for SVIs within the "medium" range for syllables in mature languages. The language delay sample showed basically similar patterning to that of the typically developing sample except that S, along with a strong negatively correlated XL, were associated with the rhythmic/syllabicity PC for the language delayed sample. The XL parameter also showed a strong positive correlation with the duration PC in the language delayed analysis. The autism sample showed simple structure, like the typically developing sample, but the nature of the structure was notably different: i) The first PC in the autism analysis was not rhythm/syllabicity (blue); ii) the spectral tilt/high pitch *a priori* parameters (green) split up and correlated with two different PCs – a high SQ correlation pertained to one and a high LT and HF to another, and ii) the PC most related to the *a priori*

wide band/low pitch grouping (violet) included a very high negative correlation for L. As in the case of the correlational results, we interpret the different autism pattern in PCA as indicating that vocalization is organized differently in terms of the acoustic parameters for the autism sample than for the other child groups. A four-factor solution was forced on the autism sample to facilitate group comparison – for a cutoff eigenvalue of one, the typically developing and language delay groups produced a four-factor PC solution, but the autism group produced a three-factor solution.

High vocal activity on the first *a priori* parameter grouping, RhSy (blue), and for its corresponding PC, can be interpreted as indicating syllable-like units are organized in the way mature syllables in speech are organized (72). The empirically determined PCA grouping (Table S10) includes VC (voicing typical of mature syllables), CS (canonical syllable formant transitions), and SE (spectral entropy typical of mature syllables), from the blue group, and M from the duration (peach) group (corresponding to duration in the medium range for mature syllables) (65). The second *a priori* parameter grouping, low spectral tilt/high pitch control (LtHp, green), includes two tilt parameters (LT, low tilt and HF, high frequency emphasis) and the SQ parameter (squeal, which indicates very high pitch, above the age-adjusted pitch range for children under four years of age or for mature syllables in speech). High correlations with these three parameters characterized a second PC in the analysis for the typically developing sample. High vocal activity for acoustic parameters in the second grouping (LtHp) suggests a vocal pattern found commonly in infants and children at very young ages, namely, high pitch along with high frequency emphasis in the spectrum (68). High vocal activity in the third grouping (BwLp), wide bandwidth (of first and second formants)/ low pitch (lower than is typical in mature speech) control, can similarly be taken to indicate a pattern often seen in playful vocalization of infancy and very young childhood. WB (wide band) and GW (growl, characterized by low pitch) were seen to have high values on a third PC, along with S (short duration) (69, 70, 68, 72). The final PC was associated in the typically developing sample with long and extra-long durations (SVIs longer than syllables in typical mature speech), again a pattern often found in very young children (65, 72).

The PCA results thus indicate that the 12 acoustic parameters pertain to relatively coherent groupings that make sense in terms of infrastructural needs for speech development. The most important grouping in terms of characterizing progression in development (at least in this implementation of the parameters) appears to have been the first one, associated with well-formed syllabification and rhythm (see below and main text, Results). Furthermore the rhythm/syllabicity grouping appears to have provided the most potent basis for differentiation of the child groups. A strong indication that the rhythm/syllabicity grouping was particularly involved can be seen in Figures S9 and S10, where the largest differences between the autism and typically developing samples occurred on the four key parameters (VC, CS, SE and M) of the first and predominant PC in the typically developing group.

The same results reported in Table S10 are displayed graphically in Figures S12a-c.

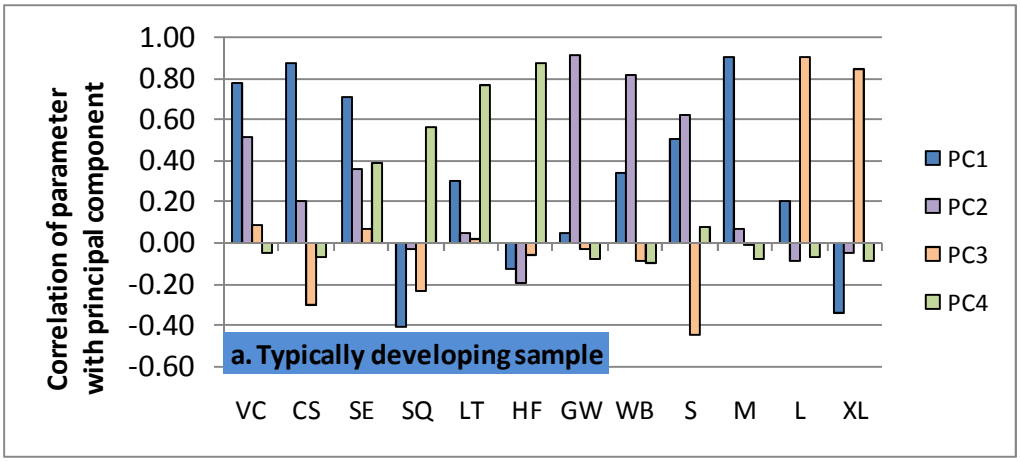

Figure S12a. PCA results for the typically developing sample at the recording level. A four-factor solution is evident (eigenvalue cutoff = 1), holding much in common with the theoretical groupings based on the infrastructural model of vocal development.

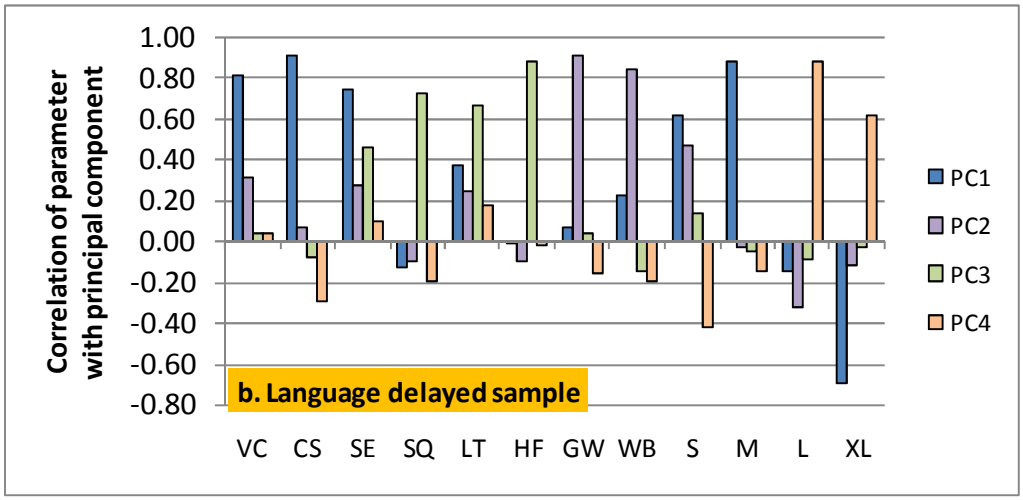

Figure S12b. PCA results for the language delay sample at the recording level. As with the typically developing sample, a four-factor solution is evident (eigenvalue cutoff = 1), holding much in common with the theoretical groupings based on the infrastructural model of vocal development.

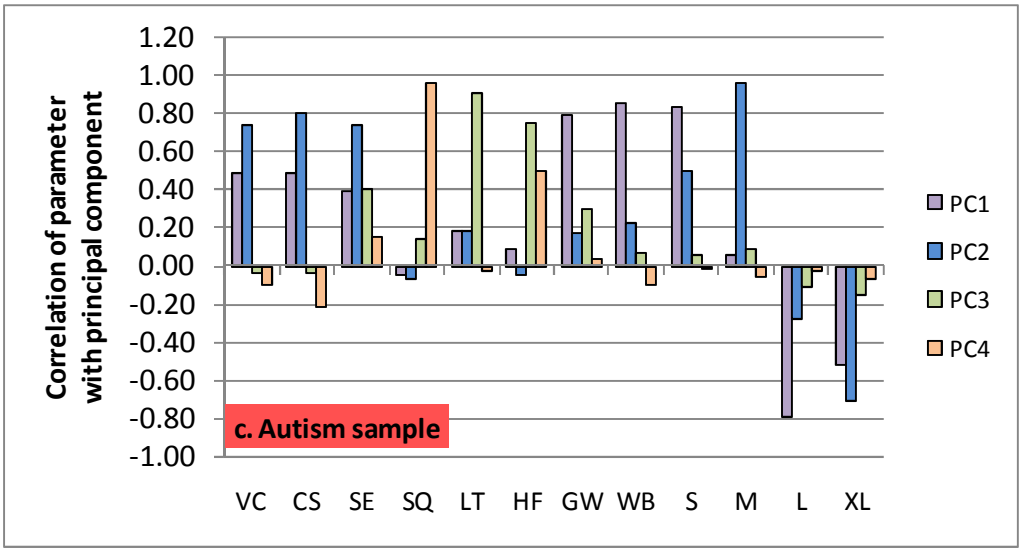

Figure S12c. PCA results for the autism sample at the recording level. The structure of the PC solution is different than in the other child groups. A four-factor solution was forced. The RhSy group is evident, but as the second PC.

Figure S13 compares mean z-scores for the three child groups on the four PCs of the typically developing sample. The figure indicates that the typically developing group used vocalization with the characteristics of the first PC, RhSy, much more often than the other child groups (note negative values on the z-scores for the other groups), suggesting that their vocalizations were more syllabically well formed than in the other groups. At the same time, the autism and language delay samples showed higher usage than the typically developing sample on the two PCs associated with pitch and spectral characteristics beyond the range that is typically found in speech. High usage of vocalizations with the acoustic characteristics of LtHp and BwLp appear to be indicative of relative immaturity of the vocal system (72). The last PC from the typically developing group (Duration) showed no significant group differences.

The group discrimination results in Figure S13 were highly statistically significant in several cases, but especially for the RhSy parameters, as can be seen in the figure. MANOVA predicting child group from the PCs of the typically developing sample showed major differences across the PCs. The RhSy PC accounted for over 21.9% of variance (based on the adjusted R^2) in group discrimination, while the three other PCs combined accounted for only about half that much (PC2 = 6.8, PC3 = 3.6 and PC4 = 1.4%). Thus the results suggest strongly that the overwhelming factor in determining group discrimination through the automated acoustic analysis of these infrastructural parameters was the extent to which children in the groups controlled well-formed rhythm/syllabification.

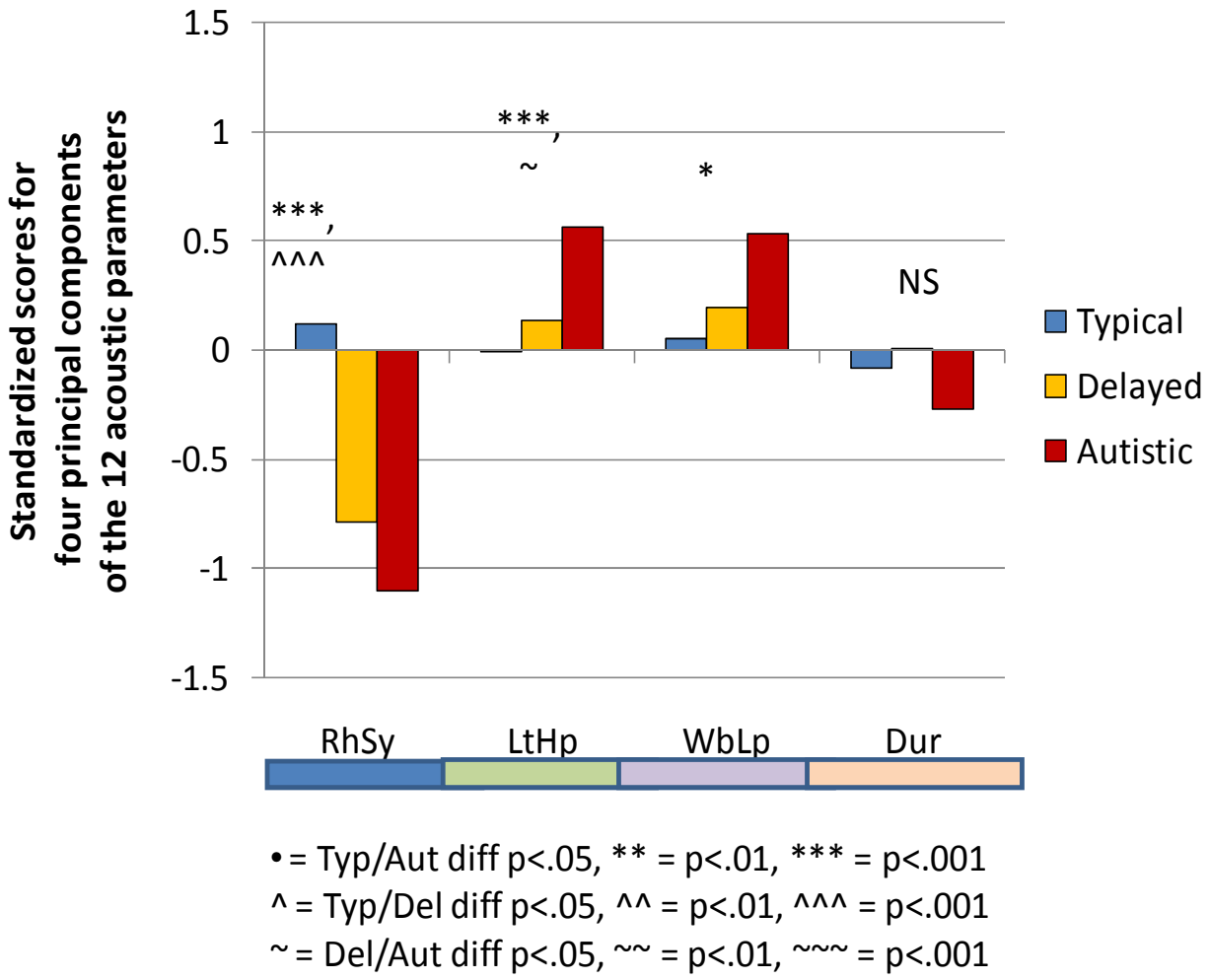

Figure S13. **Mean standardized scores (z) for four principal components compared across the three child groups**. Scores were computed based on the PCA for the typically developing sample and applied to all three groups. The scores illustrate that the rhythm/syllabicity (RhSy) grouping of parameters played a predominant role in differentiation of child groups. The typically developing sample showed higher usage of RhSy than the other groups, but lower usage on principal components associated with pitch and spectral features (LtHp, BwLp) that are unusual in mature speech.

To summarize the PCA results, they:

1. provide hopeful indications that the 12 acoustic parameters are indicative of infrastructural characteristics of speech organized into four primary groupings; and
2. suggest that parameters of syllabification and rhythm offer the most important basis for both developmental monitoring and for group differentiation among the parameters as they are currently implemented.

Individual children and subgroups of particular interest in the group discrimination analyses

In this section we explore results for certain children and subgroups of special interest. Figure S14 shows data

from the **typical vs autism** modeling configuration. The analysis with all children in the two groups can be compared with those from subsets of interest within the autism sample. For example, five children had been diagnosed originally with Pervasive Developmental Disorder (PDD) rather than with classic autism, because they were below the standard cutoff age (36 months) for diagnosis. One might have imagined these children would have been particularly hard for the modeling to detect as pertaining to the autism sample, but in fact all 5 showed posterior probabilities of autism beyond the 95% CI for the typically developing group (see second pair of bars from the left in Figure S14).

Fifteen children in the autism sample had been labeled as "echolalic", a condition common in autism, characterized by frequent repetitions of speech heard from the environment. These children might have been expected to be hard to detect as pertaining to the autism sample because they produce speech very frequently, and often it consists of well-formed syllables and other speech features found in typically developing children. The results based on the automated acoustic analysis for 13 of these children were, however, quite characteristic of the autism sample (see second pair of bars in the figure). Two of the children with echolalia, however, whose posterior probabilities are represented as a circle and a pentagon in the figure did indeed show very low probabilities, indicating that the automated procedure did not distinguish them from typically developing children.

In the MLR analysis of Figure S11c, there is an outlier (represented by very high points on the y-axis), also seen in Figure 2b (main text). This child with autism showed very high values (beyond the range for any child in any of the three groups) for the WB parameter from the BwLp grouping (reflected in Figure S11c, and perhaps as a result, on the 12 parameters together in Figure 2b). In fact the child showed an unexpected profile (clearly different from the autism group as a whole) with parameter usage that was near the end of distributions for all recordings from the three child groups on many of the 12 parameters: His recordings were high on WB, S, and CS, and very low on L, XL and SQ. He also showed relatively high values on GW, the second parameter of the BwLp grouping portrayed in Figure S11c. A non-automated acoustic evaluation (visual inspection) was conducted on a sampling of his recordings to determine if the high values for BwLp parameters might have resulted from an algorithm artifact associated, for example, with pitch-doubling. No artifact appeared to be involved. The SVIs labeled as WB and GW were indeed numerous in the recordings, and the acoustic characteristics of the examined SVIs did not show evidence of subharmonics or other spectral features that might lead to pitch-doubling in the automated analysis. In any case, it might have been imagined that this child (with his highly individual profile of parameter usage) would have been hard for the automated approach to detect as pertaining to the autism sample, but in fact his recordings showed a very high posterior probability (> 0.8) of autism (based on LDA, see red star in Figure S14).

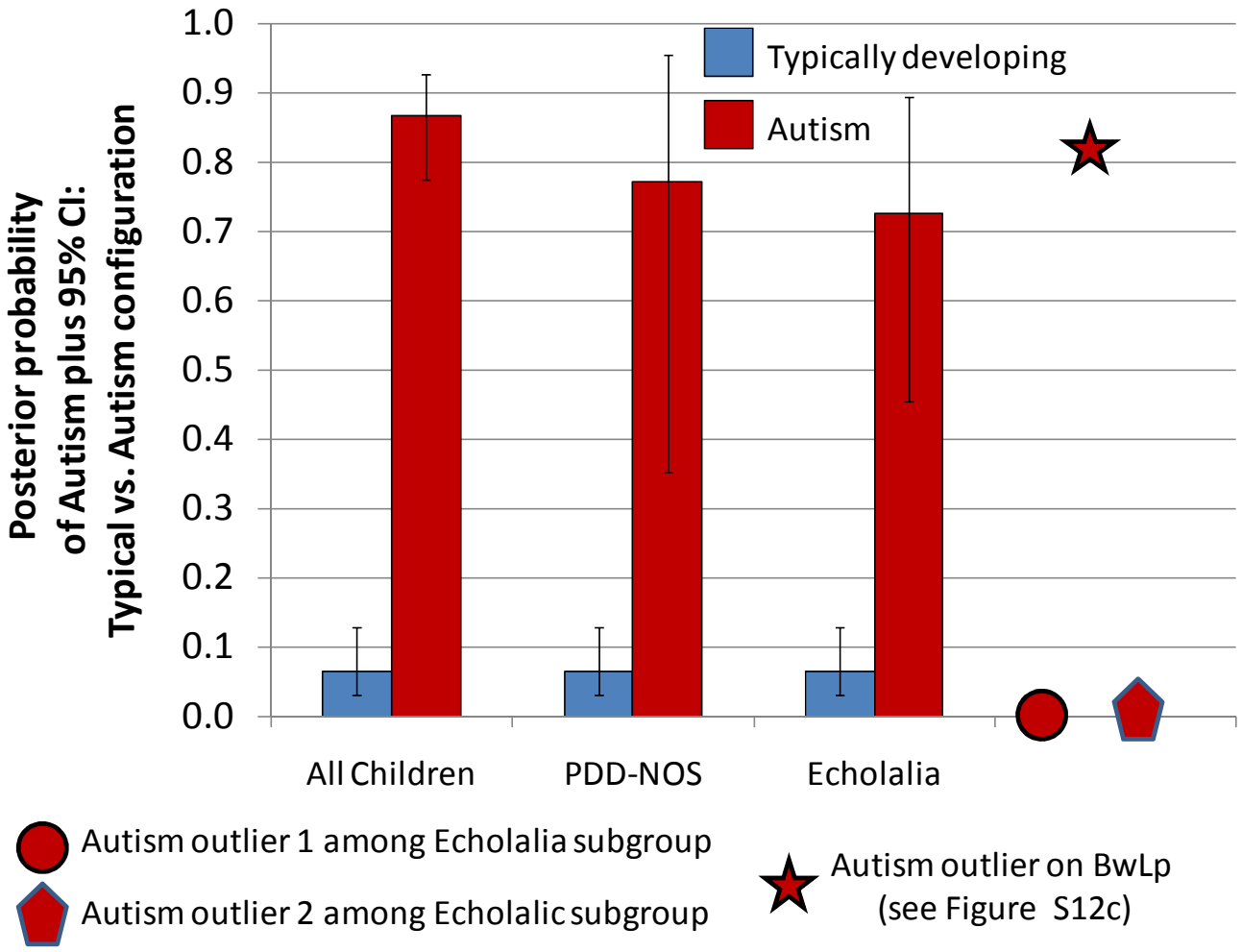

Figure S14. Results on children of special interest in group discrimination based on the modelling configuration **Typical vs Autism**. For all data in the figure, PPs were first subjected to a logit transformation, then means and CIs were computed, after which these were converted back to PPs for display; this is the same computational method used for Figure 4b (main text); without the logit transformation, the distribution of PPs is skewed with high density at high and low values; logit transformation produces a more normal distribution and allows more meaningful portrayal of means and CIs. The first pair of bars compares the entire typically developing (N = 106) and autism (N = 77) samples. Other pairs of bars show the typically developing sample against subsets from the autism sample. The second pair of bars includes data from 5 children in the autism sample originally diagnosed with pervasive developmental disorder (PDD-NOS) rather than classic autism, because they were too young at the time of diagnosis to meet requirements of classic autism. The PDD group was robustly discriminated by one-way ANOVA from the typically developing group ($p<10^{-5}$). The third pair includes 13 children with autism who were reported to have been diagnosed with echolalia. These 13 were also dramatically different from the typical sample ($p<10^{-12}$). Data for two children who were reported to have echolalia (circle and pentagon) are plotted separately as outliers, showing very low posterior probabilities (<0.01) of autism. A final child's data (star) are also plotted separately – he was the outlier in Figures S11c showing extremely high values on the BwLp parameter grouping (based on MLR) and in Figure 2b of the Main Text for all 12 parameters. Yet in spite of odd scores on BwLp (and outlying values on other parameters), his PP of autism was found to be very high (0.82, based on LDA).

Similar comparisons to those of Figure S14 are provided for the modeling configuration **typical vs autism plus language delay** in Figure S15. This configuration allows us to compare the extent to which the automated system differentiated children with (autism, language delay) and without (typically developing) language-related disorders for certain subsamples of the data. The PDD and echolalia samples show results basically similar to those of Figure S14, with highly significant differentiation from the typically developing sample ($p < 10^{-5}$). The 49 children in the language delay sample were also sharply discriminated from the typically developing sample, as illustrated in the figure ($p < 10^{-12}$). The stars in Figure S15 provide data on four of the children from the Phase I language delay sample who were designated at the time of the speech-language evaluation in Boulder as having "autistic characteristics". All of them showed very high probabilities (> 0.62, and well outside the 95% confidence interval for the typically developing sample, as portrayed in Figure S15) in the **typical vs autism plus language delay configuration**. Thus these children were designated very reliably by the automated procedure as <u>not being</u> typically developing. However, of equal interest is the fact that in the **autism vs. typical plus language delayed configuration** of modeling (not portrayed in Figures S14-15), the four children with "autistic characteristics" were also assigned very high PPs (> 0.37, well outside the range of the 95% confidence interval for PPs of a combined group of the 106 typically developing children plus the 45 language delayed children not designated as having autistic characteristics). Thus the automated procedure can be said to have assigned these four children very high probability of <u>being autistic</u>. Although we do not have direct contact with the families in most cases, we have been notified by the family that one of these four children designated as having "autistic characteristics" has recently been given a diagnosis of Asperger syndrome.

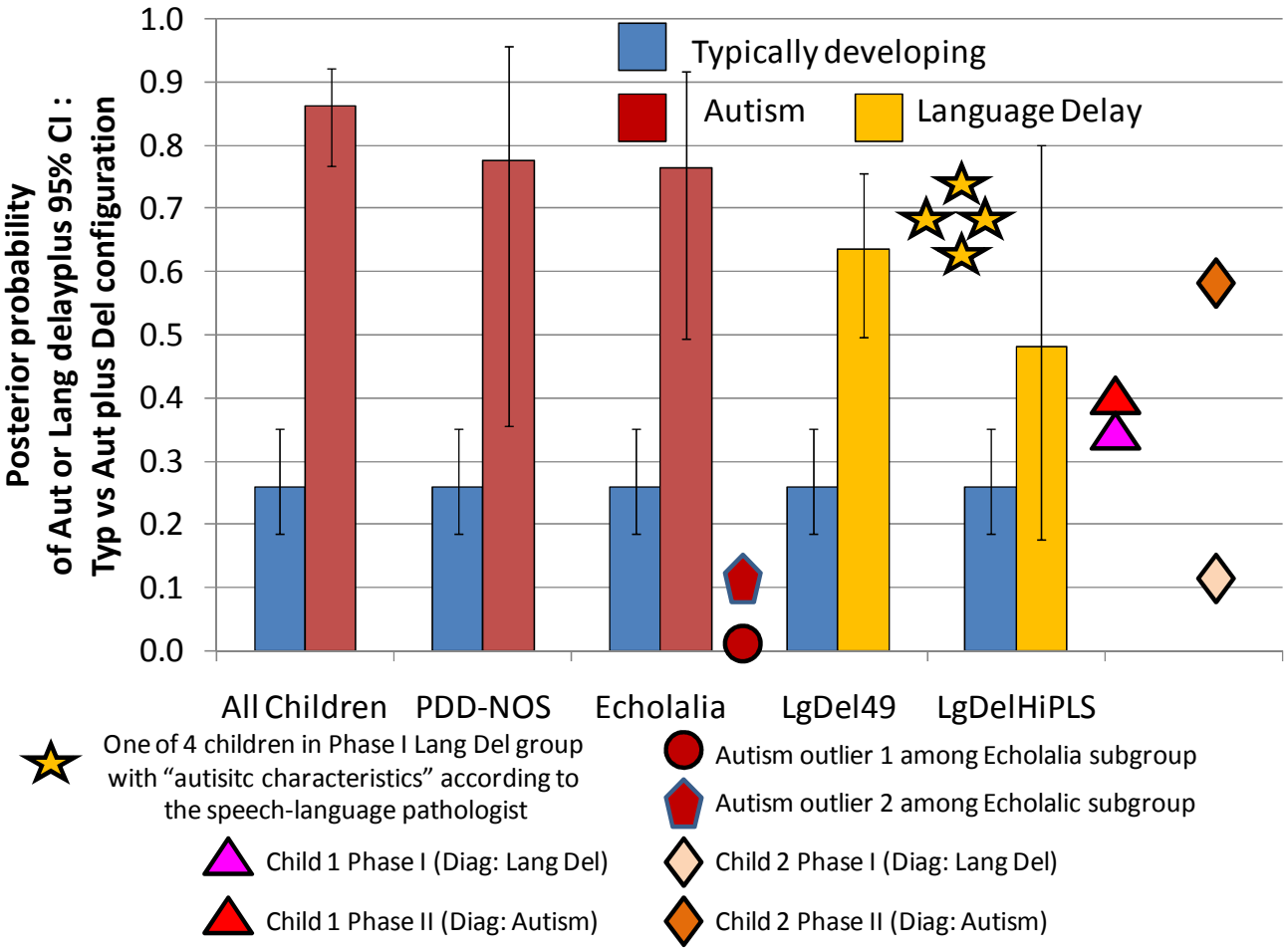

Figure S15. Results on children of special interest in the discrimination of child groups based on the modelling configuration **Typical vs Autism plus Language delay**. For computation method, see Figure S14. The first pair of bars provides a standard of comparison for typically developing (N = 106) and autism (N = 77) samples. The other pairs show outcomes for the typically developing sample against subsets from the autism sample (red) or language delayed samples (orange). The second pair of bars includes data from the PDD group, indicating the differentiation of the PDD group from the typically developing sample was highly reliable ($p<10^{-5}$) in this modelling configuration. The same observation applies to the comparison of posterior probabilities of 13 echolalic children to the probabilities for the typically developing sample ($p< 10^{-11}$). The same two outliers from the echolalic group as in Figure S14 are displayed here (circle, pentagon), again with very low posterior probabilities for this modelling configuration. The fourth pair of bars compares the typically developing sample with the entire language delay sample for this modelling configuration, and illustrates very sharp discrimination by the automated analysis ($p < 10^{-12}$). The fifth pair of bars provides data on six children from the language delay sample in Phase I who showed high scores (> 100) on the PLS-4/REEL-3 when they were tested by our speech-language pathologist months after their independent diagnosis as language delayed. These children's posterior probabilities were significantly higher than those of the typically developing sample ($p < 0.02$) as a group, but three of them had posterior probabilities within the typically developing range, suggesting (along with their high PLS/REEL scores) that they may have "caught up" with their typical peers from the point of diagnosis to the point of recording with our project. The gold stars represent four children from the language delayed sample who were designated at testing by our speech-language pathologist to have "autistic characteristics" even though none of them had to that point been diagnosed as autistic. As can be seen, all these

children were assigned very high posterior probabilities (> 0.6). One of these four has since been given a formal diagnosis of Asperger syndrome, independently of the present study. The diamonds and triangles represent the mean posterior probabilities at two different points in time for two individuals who were first assigned to the language delayed sample in Phase I (and were recorded at a mean of 15 and 25 months of age, respectively) , but whose parents later responded to the national on-line recruitment in Phase II, at which time it was reported that their diagnoses had been changed from language delayed to autistic (and then they were recorded at a mean of 33 and 44 months of age, respectively). The darker colored diamond and triangle represent the posterior probabilities for the recordings made in Phase II, and the lighter colored ones in Phase I. Because of the diagnostic change, the recordings from these two children were not included in the modelling, and they were not among either the 49 language delayed children or the 77 autistic children considered in the Main Text or prior sections of the Supporting Information Appendix. The data show one of the children having PPs a little above the mean for the typically developing sample at both Phase I and Phase II points of recording, but the other showing a very high PP at the Phase II point only, after the autism diagnosis.

Six children from the language-delayed sample in Phase I had achieved especially high mean language development scores (standard score > 100) on the PLS4 and REEL3 (the most widely used measures of early vocal communication, (51, 52)), administered by the project speech-language pathologist. These scores were of course considerably higher than expected given that the children had been previously diagnosed with language delay. The PPs for these six children displayed in Figure S15, along with their high PLS/REEL scores, suggest half of these children may have "caught up" with their typically developing peers in at least some characteristics of vocal language from the time of the diagnosis to the time of the recordings, because half of them showed PPs within the 95% CI of the typically developing sample, while the others showed PPs > 0.5.

There were two children originally involved in the Phase I language delay sample whose parents enrolled them about 20 months later through the national on-line method in the Phase II autism sample. In the interim their diagnoses had been changed by the independent professionals working with the children. Thus we had recordings on these two children at two different points in time, with differing diagnoses. We did not include these children's data in the primary samples (they were not among either the 49 language delayed or 77 autistic children considered in the Main Text or in earlier portions of the Supporting Information Appendix). However, their PPs were computed in special runs of LOOCV where their data were treated as the holdout samples for all six binary configurations of the three child groups. The results are presented as mean posterior probabilities for the configuration **Typical vs. Autism plus Language delayed** in Figure S15, and they show that Child 1 had somewhat ambiguous posterior probabilities at both Phase I and Phase II, only a little higher than the typically developing mean, while Child 2's PPs changed from being in the typical range at Phase I to being in the autistic range at Phase II. Equally pertinent are data from two modeling configurations not portrayed in Figure S15: a) **Autism vs Typical** and b) **Autism vs. Typical plus Language delayed**, where Child 1 showed posterior probabilities above the 95% CI for the typically developing sample at both Phase I and Phase II, while Child 2 fell narrowly below the typically developing means at Phase I, but showed very high posterior probabilities of autism (configuration a = 0.89, configuration b = 0.77) at Phase II.

To summarize the data on the special child cases and special subgroups:

1. Children with an early diagnosis of PDD were robustly assigned PPs in the autistic range by the automated procedure.
2. Children in the autism sample with a label of "echolalic" were robustly assigned PPs in the autistic range for 13 of 15 cases, but two were assigned PPs characteristic of typically developing children by the automated procedure.
3. The child with autism who was the primary outlier in MLR for the BwLp parameter grouping and who showed anomalous values on more than half the acoustic parameters was clearly classified as autistic by the LDA automated procedure.
4. Half the children who had been assigned to the language delay sample based on independent clinical diagnosis but who showed high (> 100) scores on language tests administered in our laboratories near the time of recordings were not identified as language-disordered by the automated procedure.
5. Children from the language delay sample who were indicated to have "autistic characteristics" by our staff speech-language pathologist, were robustly identified as being not typically developing and as being autistic as opposed to either typically developing or language delayed by the automated procedures.
6. Two special cases of children with independent professional diagnoses that changed from language delay in Phase I to autism in Phase II showed inconsistent classification by the automated procedure at different Phases and in different configurations of group comparison.

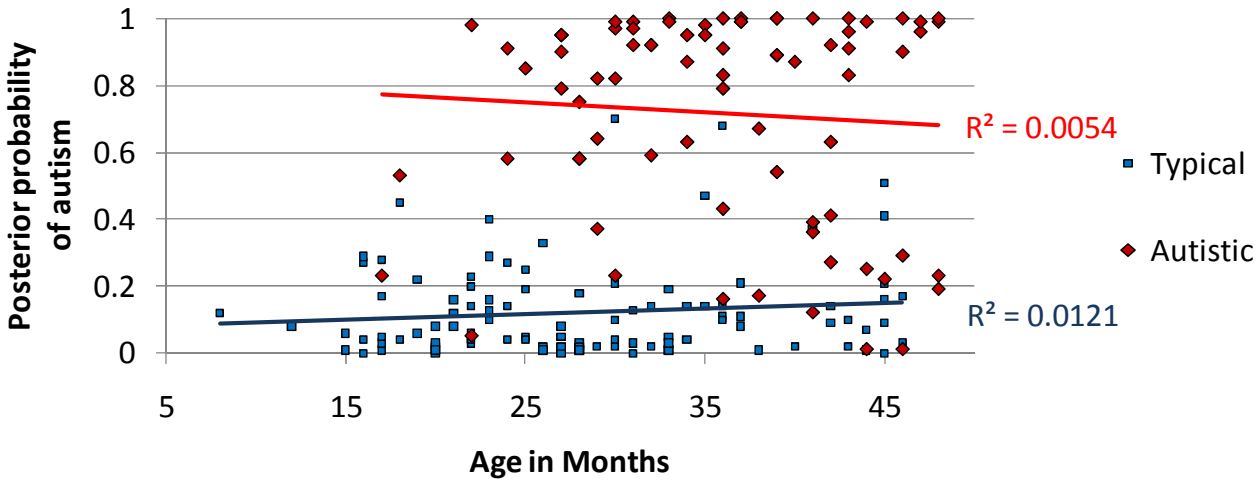

Figure S16. Correlation of posterior probability of autism with age of children in the typically developing vs. autism configuration. Extremely low correlations of PPs with age (autism r = - 0.07, typically developing r = 0.11) suggest little if any role for age in the discriminability of groups by the automated procedure for these samples.

The effect of age on group differentiation

Addressing the possibility that group differentiation by the automated method for autistic and non-autistic children might be relatively poor at the youngest ages, we correlated PPs with age for the children with autism

compared with those from the typically developing sample. The results are displayed in Figure S16. The results do not suggest any important role for age in the automated procedure's differentiation of typically developing and autism samples across the age range we have tested thus far.

SUPPORTING BACKGROUND

The problem of small sample sizes in developmental vocalization research

The study of early vocal development and its role in language has a long history. Empirical research has required laboratory recordings of children along with laborious coding, transcription, and acoustic analysis with visual inspection and measurement. Consequently sample sizes have often been very small (for reviews of relevant literature see (72, 84, 85). Representativeness of samples recorded over a half-hour to an hour (the typical durations in the laboratory) has long been viewed skeptically. Recent data illustrate that developmental change cannot be appropriately characterized in the absence of narrow-interval sampling (86).

Vocal characteristics in autism

Pronovost and colleagues (87) provided the first systematic report based on longitudinal observations of 14 children with autism, aged 5-15 years, but offered scarce quantification of their intriguing claims about anomalous voice characteristics in autism. Shriberg et al. (88) surveyed literature (including that of Pronovost and colleagues) that included commentaries or observations about prosodic characteristics of vocalizations in ASD. They found that all ten articles surveyed pointed to atypicality in the prosody of vocal affective expression, a pragmatic aspect of verbalization. Two years later an additional review (89) appeared, covering much of the same material, updating it, and reaching similar conclusions.

A sample of 15 children with ASD, aged 3-5 years, matched on nonverbal intelligence and vocal language ability with 11 children diagnosed to have other developmental delays was studied by Sheinkopf et al. (90). The research addressed vocalizations in these samples, video and audio recorded during administration of an interactive test of social and behavioral development. Auditory analysis of the recordings based on classification schemes in wide usage in infant vocalization research indicated that the ASD sample showed vocal anomalies related to pitch and voice quality in more than 20% of syllables in their utterances, more than twice as often as the delayed sample. More recent follow-ups based on the same recordings in laboratories of the first author of the present paper have scrutinized the vocalizations of both groups acoustically. Again notable differences have been found between the two groups in terms of vocal characteristics. Further acoustic work with these samples is continuing.

Research in development of vocal acoustic characteristics

Developmental acoustic changes occur in both prespeech vocalizations (babbling) and in speech (91-93). Of primary interest here are developments that pertain to *both* speech and prespeech, and thus can be

monitored starting in the first months of life and followed across years. These can be termed "infrastructural" (or "infraphonological") (72), because they manifest elements of vocal control required for any spoken language. For example, all languages use normal phonation (i.e., voicing) as a source of acoustic energy for the production of speech events. Normal phonation has acoustic consequences that can be monitored from the first months of life and contrasted with other types of phonation (68, 83, 94). In addition all languages involve supraglottal articulatory actions that filter or modulate the phonatory stream during speech events. These modulations are produced rhythmically, creating acoustically identifiable syllables as minimal rhythmic units in languages. Supraglottal modulations can also be monitored systematically from the first months of life (95). Finally, the syllable-producing articulatory modulations of speech entail characteristic cyclic durations. Thus, syllables tend to have durations within a fixed range, and this durational pattern can also be monitored starting from the first months of life (96).

The degree of infant control over these infrastructural properties of speech (phonation, syllabicity, and duration of syllabic units) can be monitored without direct reference to the concrete phonological elements that constitute the mature system of speech segments, the phonemes. The lack of necessary reference to phonemes in the acoustic analysis we used is important because infants in the first months of life do not command phonemes at all in vocal production. Further, it is now widely believed by child phonologists that truly phonemic control in speech production is not achieved until at soonest the second year of life (80, 85). And when a truly phonemic system does begin to emerge, its contrastive elements are fewer and more variable in form than those of the mature system. So instead of beginning with phonemic development, infants begin life by building infrastructural capabilities such as phonatory control, along with the systematic articulatory movements that are required for syllables (97, 98).

Parameter groupings a-c (Table S5) all include parameters that monitor phonatory action and its development – utterances with very high or very low pitch of course represent substantial variations in phonatory properties. These groupings were selected in part to reflect findings of research in infant vocal development, which has repeatedly reported an early three-way contrast (proposed to be the earliest voluntary vocal contrast of human infancy) among utterances that are described in the terminology of the field as vowel-like sounds (or vocants), squeals, and growls (79, 99, 100). Vowel-like sounds include phonation produced in a mid-pitch range for the particular voice in question, squeals provide a high-pitched contrast with vowel-like sounds, and growls provide a low-pitched contrast with vowel-like sounds. Parameter grouping d was also selected to reflect a documented tendency for systematic change with age, in this case regarding durations of vocalizations (101).

Because of the fact that our automated analysis was conducted on SVIs, roughly syllabic-like units, the parameters of grouping a (RhSy) plus the medium duration parameter appear to hang together. They appear to

provide a basis for monitoring the emergence of well-formed syllabification, because each of the four parameters monitors a property of well-formed syllables (voicing, canonical transitions, spectral entropy within the expected range for syllables, and duration typical of mature syllables).

Automated acoustic analysis

The great bulk of research that has been conducted in infant vocalizations and child phonology has been based on auditory analyses, usually including phonetic transcription (102-104), although a number studies have included acoustic analysis as well (105, 106). Automated analysis of vocalizations in infancy and early childhood has been very rare. One reason is that early infant vocalizations do not contain phonemic elements, and consequently, the primary available methods of automatic speech recognition (which are very predominantly based on phonemic principles) are seemingly inapplicable.

One important effort at automated categorization of acoustic features in infant vocalizations has focused on canonical syllables and related phenomena (107). The approach has been based primarily on automated evaluation of acoustic landmarks (108) associated with onsets of syllables and has targeted differentiation of infants and children with and without communication disorders. The work has achieved some success in identifying canonical syllables in infancy and in differentiating groups, and provides hope that automated methods can do much more in the future. In addition, another project centered at the Oregon Health and Science University is underway under funding from NIDCD to evaluate prosody in autism with automated tools and a preliminary report provides reason for optimism about the approach (109).

From the standpoint of the goals of the present work, a critical limitation of the prior efforts (both of Fell and colleagues and of the Oregon Health and Science University group) is that the approaches have not been applied directly to naturalistic recordings. Instead, the efforts either require that listeners precode data from recordings, locating utterances by visual/auditory inspection, or the utterances are the product of a specific laboratory/clinical elicitation task, where the target word or words are known for the speaker prior to automated analysis. Thus particular utterances deemed appropriate for analysis are preselected rather than being found by the automated analysis within naturally occurring vocal communication. These approaches do not have to differentiate babbling and speech from cries and vegetative sounds, vocalizations produced by other speakers, or other irrelevant noises. Precoding by human observers and conducting elicitation tasks with human experimenters in the laboratory or clinic are time consuming and costly (although the Oregon Health and Science University group is developing computer interactive elicitation procedures for some of its measures). The data analyzed in such approaches is not as naturalistic as the data analyzed here, and thus the method is subject to question with regard to representativeness of the child's actual performance in communication. For the present goals, such precoding or elicitation would defeat the primary purpose – we seek to provide a fully

automated procedure, where all the steps of the analysis are included starting with a raw, naturalistically acquired acoustic waveform.

The need for interdisciplinary cooperation in this research

A key problem in establishing productive relations between engineering and vocal development for the purposes of automated acoustic analysis is to establish a common set of goals. In general, speech recognition engineers work from models that begin with mature phonemics as an organizing principle – thus, for example, Gaussian mixture models in current speech recognition software typically target individual phonemes (i.e., alphabetical level consonant or vowel segments, /t/ , /n/, /i/, etc.) or sequences of phonemes. In contrast, research in infant vocalizations has long rejected the idea that infants in the first months of life command phonemes at all, both on theoretical and empirical grounds. Further, across the first several years of life, phonemic control emerges in stages and degrees – there is no fixed point of phonemic onset. This perspective suggests that an optimal approach to acoustic modeling across the first years should target *infrastructural* properties of vocal development that show relatively continuous growth across several years. Hence the marriage established for the present work involved engineering to implement infrastructural properties (rather than phoneme-based models) in the form of 12 parameters that were designed on the basis of theoretical considerations current in research on infant vocalizations and young child phonology.

SUPPORTING REFERENCES

45. Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J. & Hansen, J. (2008) Signal Processing for Young Child Speech Language Development, in *The 1st Workshop on Child, Computer and Interaction*, Chania, Crete, Greece).
46. Zimmerman, F., Gilkerson, J., Richards, J., Christakis, D., Xu, D., Gray, S. & Yapanel, U. (2009) Teaching By Listening: The Importance of Adult-Child Conversations to Language Development. *Pediatrics* **124,** 342-349.
47. Christakis, D., Gilkerson, J., Richards, J., Zimmerman, F., Garrison, M., Xu, D., Gray, S. & Yapanel, U. (2009) Audible TV is associated with decreased adult words, infant vocalizations, and conversational turns: A population based study. *Archives of Pediatrics and Adolescent Medicine* **163,** 554-558.
48. Warren, S. F., Gilkerson, J., Richards, J. A. & Oller, D. K. (2010) What Automated Vocal Analysis Reveals About the Language Learning Environment of Young Children with Autism. *Journal of Autism and Developmental Disorders* **40**, 555-569.
49. Gilkerson, J. & Richards, J. A. (2008) in *Infoture Technical Reports, ITR-02-2* (Infoture Technical Reports, ITR-02-2, Boulder, CO).
50. Gilkerson, J. & Richards, J. A. (2008) in *Infoture Technical Reports, ITR-07-2* (Infoture Technical Reports, ITR-07-2, Boulder, CO).
51. Zimmerman, I. L., Steiner, V. G. & Pond, R. E. (2008) *Preschool Language Scale, Fourth Edition (PLS-4) English Edition*. (Pearson Education Inc., San Antonio, TX).
52. Bzoch, K. R., League, R. & Brown, V. L. (2006) *Receptive-Expressive Emergent Language Test-Third Edition (REEL–3)*. (Pearson Education Inc., San Antonio, TX).
53. Ireton, H. & Glascoe, F. P. (1995) Assessing children's development using parents' reports. The Child Development Inventory. *Clinical Pediatrics***,** 248-55.
54. Achenbach, T. M. (1991) *Integrative Guide to the 1991 CBCL/4-18, YSR, and TRF Profiles*. (University of Vermont Department of Psychiatry, Burlington, VT).
55. Robins, D. L., Fein, D., Barton, M. L. & Green, J. A. (2001) The Modified checklist for autism in toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. *Journal of Autism and Developmental Disorders* **31,** 131-44.
56. Wetherby, A. M. & Prizant, B. M. (2001) *Communication and Symbolic Behavior Scales Developmental Profile™ (CSBS DP™)*. (Brookes, Baltimore, MD).
57. Sikora, D. M., Hall, T. A., Hartley, S. L., Gerrard-Morris, A. E. & Cagle, S. (2008) Does Parent Report of Behavior Differ Across ADOS-G Classifications: Analysis of Scores from the CBCL and GARS. *Journal of Autism and Developmental Disorders* **38,** 440-448.
58. Watt, N., Wetherby, A. M. & Barber, A. (2008) Repetitive and Stereotyped Behaviors in Children with Autism Spectrum Disorders in the Second Year of Life. *Journal of Autism and Developmental Disorders* **38,** 1518-1533.
59. Lehiste, I. (1973) Rhythmic units and syntactic units in production and perception. *Journal of the Acoustical Society of America* **54,** 1228-1234.
60. Nakatani, L. H., O'Conner, K. D. & Aston, C. H. (1981) Prosodic aspects of American English speech rhythm. *Phonetica* **38,** 84-106.
61. Cruttenden, A. (1986) *Intonation*. (Cambridge University Press, New York).
62. Dechert, H. W. & Raupach, M. (1980) *Temporal variables in speech: Studies in honor of Frieda Goldman-Eisler*. (Mouton, The Hague).
63. Oller, D. K. & Lynch, M. P. (1992) Infant vocalizations and innovations in infraphonology: Toward a broader theory of development and disorders. in *Phonological Development: Models, Research, Implications*, eds. Ferguson, C., Menn, L. & Stoel-Gammon, C. (York Press, Inc., Parkton, MD), pp. 509-538.
64. Lynch, M. P., Oller, D. K., Steffens, M. L. & Buder, E. H. (1995) Phrasing in prelinguistic vocalizations. *Developmental Psychobiology* **28,** 3-23.

65. Nathani, S. (1998) Doctoral dissertation in *Audiology and Speech Sciences* (Purdue University, West Lafayette, IN).
66. Rochester, S. R. (1973) The significance of pauses in spontaneous speech. *Journal of psycholinguistic research* **2,** 51-81.
67. Delattre, P. (1965) *Comparing the phonetic features of English, French, German and Spanish*. (Julius Groos Verlag, Heidelberg).
68. Stark, R. E. (1981) Infant vocalization: A comprehensive view. *Infant Medical Health Journal* **2,** 118-128.
69. van der Stelt, J. M. (1993) *Finally a word: a sensori-motor approach of the mother-infant system in its development towards speech*. (Uitgave IFOTT, Amsterdam, the Netherlands).
70. Roug, L., Landberg, I. & Lundberg, L. J. (1989) Phonetic development in early infancy: a study of four Swedish children during the first eighteenth months of life. *Journal of Child Language* **16,** 19-40.
71. Oller, D. K. (1986) Metaphonology and infant vocalizations. in *Precursors of early speech*, eds. Lindblom, B. & Zetterstrom, R. (Stockton Press, New York), pp. 21-35.
72. Oller, D. K. (2000) *The Emergence of the Speech Capacity*. (Lawrence Erlbaum Associates, Mahwah, NJ).
73. Oller, D. K. & Griebel, U. (2008) Complexity and flexibility in infant vocal development and the earliest steps in the evolution of language. in *Evolution of Communicative Flexibility: Complexity, Creativity and Adaptability in Human and Animal Communication*, eds. Oller, D. K. & Griebel, U. (MIT Press, Cambridge, MA), pp. 141-168.
74. Oller, D. K., Wieman, L., Doyle, W. & Ross, C. (1975) Infant babbling and speech. *Journal of Child Language* **3,** 1-11.
75. Oller, D. K. & Eilers, R. E. (1975) Phonetic expectation and transcription validity. *Phonetica* **31,** 288-304.
76. Oller, D. K. & Ramsdell, H. L. (2006) A weighted reliability measure for phonetic transcription. *Journal of Speech and Hearing Research* **49,** 1391-1411.
77. Ramsdell, H. & Oller, D. K. (2007) Predicting phonetic transcription agreement: Insights from research in infant vocalizations. *Clinical Linguistics and Phonetics* **21,** 793-831.
78. Feingold, M. (1992) The Equivalence of Cohen's Kappa and Pearson's Chi-Square Statistics in the 2 X 2 Table. *Educational and Psychological Measurement* **52,** 57-61.
79. Oller, D. K. (1980) The emergence of the sounds of speech in infancy. in *Child phonology, Vol 1: Production*, eds. Yeni-Komshian, G., Kavanagh, J. & Ferguson, C. (Academic Press, New York), pp. 93-112.
80. MacNeilage, P. F. & Davis, B. L. (1990) Acquisition of speech production: The achievement of segmental independence. in *Speech production and speech modelling*, eds. Hardcastel, W. J. & Marchal, A. (Kluwer, Dordrecht), pp. 55-68.
81. Zemlin, W. R. (1998) *Speech and Hearing Science Anatomy and Physiology*. (Allyn and Bacon, Needham Heights, MA).
82. Fairbanks, G. (1959) *Voice and articulation drillbook*. (Harper and Row, New York).
83. Buder, E. H., Chorna, L., Oller, D. K. & Robinson, R. (2008) Vibratory Regime Classification of Infant Phonation. *Journal of Voice* **22,** 553-564.
84. Locke, J. L. (1993) *The child's path to spoken language*. (Harvard University Press, Cambridge, Massachusetts).
85. Vihman, M. M. (1996) *Phonological Development: The Origins of Language in the Child*. (Blackwell Publishers, Cambridge, MA).
86. Adolph, K. E., Robinson, S. R., Young, J. W. & Gill-Alvarez, F. (2008) What is the shape of developmental change? *Psychological Review* **115,** 527-543.
87. Pronovost, W., Wakstein, M. P. & Wakstein, D. J. (1966) Longitudinal study of the speech behavior and language comprehension of fourteen children diagnosed atypical or autistic. *Exceptional Children* **33,** 19-26.
88. Shriberg, L. D., Paul, R., McSweeny, J. L., Klin, A., Cohen, D. J. & Volkmar, F. R. (2001) Speech and prosody characteristics of adolescents and adults with high functioning autism and asperger syndrome. *Journal of Speech, Language, and Hearing Research* **44,** 1097-1115.

89. McCann, J. & Peppe, S. (2003) Prosody in autism spectrum disorders: a critical review. *International Journal of Language and Communication Disorders* **38,** 325-350.
90. Sheinkopf, S. J., Mundy, P., Oller, D. K. & Steffens, M. (2000) Vocal atypicalities of preverbal autistic children. *Journal of Autism and Developmental Disorders* **30,** 345-353.
91. Kent, R. D. (1981) Articulatory-acoustic perspectives on speech development. in *Language Behavior in Infancy and Early Childhood*, ed. Stark, R. (Elsevier, New York), pp. 105-126.
92. Koopmans-van Beinum, F. J. & van der Stelt, J. M. (1986) Early stages in the development of speech movements. in *Precursors of early speech*, eds. Lindblom, B. & Zetterstrom, R. (Stockton Press., New York), pp. 37 50.
93. Robb, M. P., Chen, Y. & Gilbert, H. R. (1997) Developmental aspects of formant frequency and bandwidth in infants and toddlers. *Folia Phoniatrica et Logopaedica* **49,** 88-95.
94. Laufer, M. Z. & Horii, Y. (1977) Fundamental frequency characteristics of infant non-distress vocalizations during the first twenty-four weeks. *Journal of Child Language* **4,** 171-184.
95. Oller, D. K. (1981) Infant vocalizations: Exploration and reflexivity. in *Language behavior in infancy and early childhood*, ed. R.E.Stark (Elsevier North Holland, New York), pp. 85-104.
96. Robb, M. P. & Saxman, J. H. (1990) Syllable durations of preword and early word vocalizations. *Journal of Speech and Hearing Research* **33,** 583-593.
97. Holmgren, K., Lindblom, B., Aurelius, G., Jalling, B. & Zetterstrom, R. (1986) On the phonetics of infant vocalization. in *Precursors of early speech*, eds. Lindblom, B. & Zetterstrom, R. (Stockton Press., New York), pp. 51 63.
98. Oller, D. K. & Eilers, R. E. (1992) Development of vocal signaling in human infants: Toward a methodology for cross-species vocalization comparisons. in *Nonverbal vocal communication*, eds. Papoušek, H., Jürgens, U. & Papoušek, M. (Cambridge University Press, New York), pp. 174-191.
99. Stark, R. E. (1980) Stages of speech development in the first year of life. in *Child Phonology, vol. 1*, eds. Yeni-Komshian, G., Kavanagh, J. & Ferguson, C. (Academic Press, New York), pp. 73-90.
100. Nathani, S., Ertmer, D. J. & Stark, R. E. (2006) Assessing vocal development in infants and toddlers. *Clinical Linguistics and Phonetics* **20,** 351-367.
101. Hsu, H. C., Fogel, A. & Cooper, R. B. (2000) Infant vocal development during the first 6 months: Speech quality and melodic complexity. *Infant & Child Development* **9,** 1-16.
102. Locke, J. L. (1983) *Phonological acquisition and change*. (Academic Press, New York).
103. Ingram, D. (1989) *First language acquisition: Method, description, and explanation*. (Cambridge University Press, New York).
104. Stoel-Gammon, C. (1992) Research on phonological development: Recent advances. in *Phonological development models, research, implications*, eds. Ferguson, C. A., Menn, L. & Stoel-Gammon, C. (York press, Timonium, MD), pp. 273-282.
105. Buder, E. H. & Stoel-Gammon, C. (2002) Young children's acquisition of vowel duration as influenced by language: Tense/lax and final stop consonant voicing effects. *Journal of the Acoustical Society of America* **111,** 1854-1864.
106. Sussman, H. M., Duder, C., Dalston, E. & Cacciatore, A. (1999) An acoustic analysis of the development of CV coarticulation: A case study. *Journal of Speech, Language, and Hearing Research* **42,** 1080-1096.
107. Fell, H. J., MacAuslan, J., Ferrier, L. J. & Chenausky, K. (1999) Automatic babble recognition for early detection of speech related disorders. *Journal of Behaviour and Information Technology* **18,** 56-63.
108. Stevens, K. N. (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America* **111,** 1872-1891.
109. Prud'hommeaux, E. T., van Santen, J., Paul, R. & Black, L. (2008) (International Meeting for Autism Research.